

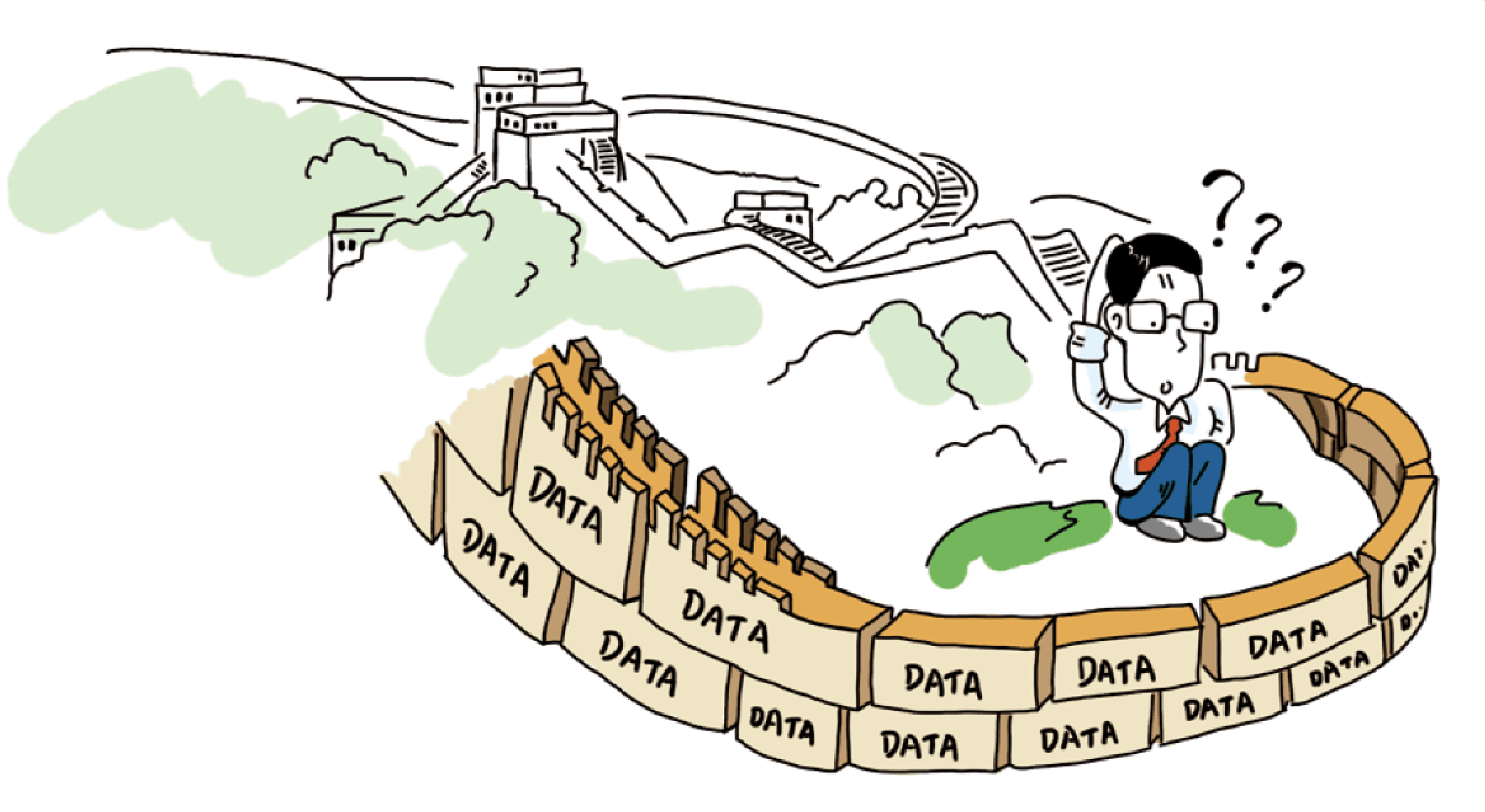
让世界更清晰

掌握
数据爆炸时代
先人一步的
新思维

大话数据挖掘

西安美林电子有限责任公司 编著

清华大学出版社



大话数据挖掘

西安美林电子有限责任公司　编著

清华大学出版社
北　京

内 容 简 介

本书以 EMBA 班的“数据挖掘技术及其应用”教学为场景，带领读者步入数据挖掘的神秘殿堂，领略数据挖掘的神奇魅力。全书分为 9 章：第 1 章从三个真实故事开始数据挖掘之旅；第 2 章以某企业生产中遇到的质量控制难题的解决过程为线索，展现数据挖掘的实施过程；第 3 章到第 9 章以典型案例的形式分别介绍了数据挖掘技术在电力行业、交通航空领域、冶金行业、税务与金融行业、电信行业、故障诊断以及互联网行业的应用。

数据挖掘是一种专业性极强的技术，本书避开大量晦涩的概念和令人生畏的数学公式，以师生互动讨论的形式让读者走进数据挖掘殿堂，进而深入浅出、循序渐进地感知数据挖掘。随着阅读，读者会自然而然地身临课堂，“让数据说话，从数据中发现规律，科学决策”等新的理念会使读者对实际工作中面临的复杂问题浮想联翩、另辟新径。

本书适合企事业部门的领导、管理人员、生产一线的技术人员，另外，学生或者行业工作者，可以通过本书的阅读，为以后的学习奠定好基础。

图书在版编目(CIP)数据

大话数据挖掘 / 西安美林电子有限责任公司编著. —北京：清华大学出版社，2012

ISBN 978-7-302-29813-7

Ⅰ.①大…　Ⅱ. ①西…　Ⅲ. ①数据采集　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2012)第 190096 号

责任编辑：栾大成
封面设计：
责任校对：徐俊伟
责任印制：
出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，　http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn

印 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：185mm×230mm　**印　　张**：17.75　**插页**：1　**字　　数**：343 千字
版　　次：2013 年 1 月第 1 版　**印　　次**：2013 年 1 月第 1 次印刷
印　　数：1~5000
定　　价：39.00 元

产品编号：048928-01

前　言

本书的萌发

上世纪 80 年代末到 90 年代初，国内外广泛流传着一句耐人寻味的话语：**我们沉浸在数据的海洋中，却渴望着知识的淡水**。这句话生动地描绘了当时人们面对海量数据的迷惘和无奈。就在这时，世界商业巨头沃尔玛从其庞大的交易数据库中演绎了一场“啤酒和尿布的故事”，揭示了一条隐藏在海量数据中的、美国人的一种行为规律：年龄在25～35岁的年轻父亲下班后经常要到超市去给婴儿买尿布，而他们中有30%～40%的人顺手为自己买几瓶啤酒。受这条简单的客户行为模式的启发，沃尔玛调整了商品布局，并策划了促销价格，结果销售量大增。这一现象引起了科学界的注意，他们将“啤酒和尿布的故事”引申为“关联规则获取”，进而将**“从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又潜在有用的信息和知识的过程”**定义为**“数据挖掘”**。

需求是成功之源，于是西方发达国家刮起了一场数据挖掘的风暴。商业界发现了沃尔玛迅猛发展的秘诀，纷纷效仿。电信行业也沸腾了，各公司纷纷争先恐后地利用数据挖掘这一锐利武器解决他们面临的最紧迫的问题（如客户分群、客户会流失原因及预测、业务套餐及响应、关联消费等）。工业界也行动了，他们从堆积如山的数据中，挖掘出指导生产和管理的决策规则。

上世纪 90 年代中期以后，基于数理统计、人工智能、机器学习、人工神经网络等多种技术的数据挖掘技术已经成为研究和应用的热点，数据挖掘在我国也开始推广应用。然而，从这么多年的情况来看，我国数据挖掘的应用与发达国家还有很大差距。我们仅在互联网、金融、电信和商业等领域有一些成功的应用，而在其他行业如制造、航空、医药、军工、化工、税务、反恐和刑侦等只有少量的尝试。为什么会这样呢？IT 界、企业界和学术界的有识之士无不在思考着这样的问题。进行数据挖掘，数据是基础，难道是我国的信息化建设还未达到一定的程度，数据积累不够？

进入 21 世纪前可以这么说，可现在，显然不是。目前，我国的大中型企业，大多建立了先进的信息化系统，甚至相当多的企业构建了数据仓库，而且数据日复一日、爆炸式地增长，可谓堆积如山。然而，很多企业对数据挖掘的认识还不全面，甚至感

觉其神秘不可信，这样的话，生产管理中遇到了不能解决的问题，自然不会用数据挖掘的思想思考，甚至基层部门提出使用这样的方法，管理层却因对此不甚了解而无力推动。

为此，我们期望从领导层和生产一线的工作人员普及数据挖掘知识开始，唤起人们对数据新的认识：数据是客观实际的反映，它体现了营销规律、生产规律、经营规律和产品质量控制规律。更重要的是，使企业管理告别基于简单统计分析的“报表”决策时期，跨入数据挖掘的“知识”决策时代。

为了实现这一目标，迫切地需要一本使企业管理者和基层工作者喜闻乐见的读物。然而，市面上的数据挖掘书籍几乎全是教科书形式，理论性太强，满篇数学公式，让人望而却步，而且应用实例甚少，让人难以理解。在这种情况下，我们大胆地萌发出一种案例教学法编写思路，以课堂教学为线索，介绍数据挖掘的基本概念和应用过程，让读者轻松地走进数据挖掘，领略数据挖掘的神奇魅力。

本书的读者群

如果您是一位企业或政府部门的领导，您可以利用乘飞机的闲暇，与本书中的徐教授和各行各业的精英们一起，走进数据挖掘的世界，相信当您下飞机的时候，一定会浮想联翩，产生许多新的思路；

如果您是一位企业管理、生产一线的技术人员，利用一个周末的休息时间，通过本书，您会对数据挖掘有初步而较为系统的了解和认识，您会自觉地尝试利用数据挖掘的方法解决实际问题；

如果您是一位想系统学习数据挖掘知识的学生或科技工作者，亦可以通过本书的阅读，为以后的学习奠定好基础。

本书的内容

全书共 9 章。第 1 章，揭开数据挖掘的面纱，从三个真实而有趣的故事开始，让读者了解数据挖掘的概念、数据挖掘产生与发展、数据挖掘的功能和数据挖掘技术，本章深入浅出地介绍了关联规则、聚类分析、预测（分类和回归）、时间序列等数据挖掘方法及常用算法；第 2 章简述数据挖掘流程，以某冶金企业生产中遇到的质量控制技术攻关难题的解决过程为线索，活灵活现地展现了一个数据挖掘问题的项目立项

及其实施过程；第 3 章到第 9 章以典型案例的形式分别介绍了数据挖掘技术在电力行业、交通航空领域、冶金行业、税务与金融行业、故障诊断、电信行业、互联网行业方面的应用。

本书的特色

形式新颖

本书以 EMBA 班的“数据挖掘技术及其应用”教学为场景，通过教师与学员互动共鸣的形式，带领读者步入数据挖掘的神秘殿堂，领略数据挖掘的神奇魅力。这种写作方式，避免了传统教科书理论性太强，数学公式繁多，让非专业数据挖掘者望而却步的缺陷。

案例导读

本书通过数据挖掘的典型案例，引导读者领略如何利用数据挖掘技术解决各行各业生产和管理中的实际问题。摈弃了晦涩难懂的理论，在解决问题的过程中了解数据挖掘技术及其应用方法，学会**“让数据说话，以数据辅助决策”**的新理念。

创作团队

本书由西安交大美林数据挖掘研究中心策划，靖稳峰、卢耀宗等编写，程宏亮为本书审定了章节划分并精选了案例素材，王璐为本书审定了故事构思和语言风格，程宏斌、李炜、强劲和黄蓉等对本书提出了大量的建设性构想和修改意见，并参与了部分章节的编写。陈浩铭和王羽为本书制作了精美插图。

致谢

西安交通大学徐宗本院士在百忙中对本书的构思、写作给予了悉心指导，清华大学出版社栾大成编辑对本书原稿字斟句酌，使得本书增色不少，这里一并表示衷心感谢。西安交大美林数据挖掘研究中心还有许多同事为本书的出版付出了大量心血，在此表示诚挚的谢意。

编者

目　录

第 1 章　揭开数据挖掘的面纱

徐教授是某985院校的著名教授，国内数据挖掘专家、智能信息处理研究方向学术带头人，主持了20多项国家项目和国际合作项目，具有丰富的数据挖掘项目实施经验，获得过多项国家级大奖。数十年来，他潜心科研，除了给自己学院的本科生和研究生上课外，一直谢绝其他授课邀请。这次他破例了，欣然接受了本校管理学院第5届EMBA班的“数据挖掘及其应用”课程……



1.1 历史的使命

今天是第一节课，徐教授一跨进教室，迎接他的是学员们一阵热烈的掌声。他习惯性地扫视了一下学生，果然正像管理学院张院长介绍的那样，在座的学员不同寻常，年龄在35～50岁之间，个个西装革履，精神焕发，眼睛里放射出对新知识无比渴望的光芒。

徐教授走上讲台，先在黑板上写下了自己的名字和联系方式，然后微笑着说："同学们，今天我能站在这儿给大家上课，不是因为你们管院张院长有面子，也不是因为你们这些学员地位有多高，说实在的，是党中央、国务院让我来的。"学员们个个目瞪口呆。

有人嘀咕道："难道中央还关心我们这个EMBA班？。"

"关心，而且非常关心。"徐教授铿锵有力地回答。

大家更加疑惑了。

徐教授提高了嗓门："2006年1月9日，在全国科技大会上，党中央、国务院作出了建设创新型国家的重大决策。大家都知道，创新型国家是指以技术创新为经济社会发展核心驱动力的国家。技术创新需要科学家和科技工作者的努力，更离不开政府和企业高层领导和管理人员的推动。张院长在邀请我来给你们上课时介绍说，在座各位都在政府部门或者企业地位显赫，所以我欣然地、破天荒地答应了你们院长的邀请。不过，别以为是你们的乌纱帽吸引了我，而是你们每一个人身上肩负的'建设创新型国家'的历史使命召唤着我。"

徐教授越说越激动，喝了口水继续说："我为科学事业奋斗了一辈子，深知'象牙塔'里的发明、创造，需要与经济建设结合才更能体现出其价值，才更能为建设创新型国家做出贡献。理论创新的成果要真正转化为生产力，迫切需要一种推动力、催

化剂。而能起到这种作用的主体非你们这些人莫属，诚如是，你们就是建设创新型国家的排头兵。你们说，党中央能不关心你们吗？”



徐教授的话音刚落，教室里立刻响起长时间的掌声。

他双手从上向下慢慢挥动，示意大家停下，接着说：“近十年来数据挖掘技术飞速发展，在国外，数据挖掘正在变成整个信息技术的核心之一。尤其是世界500强企业均设立了数据挖掘研发与应用部门，数据挖掘技术已成为其业务成功的关键因素。2007年5月，《纽约时报》以‘**数据挖掘正在进入主流**’为题，介绍了数据挖掘技术，并指出这种新技术正在变成人们工作和生活中不可或缺的一个部分。”

徐教授停顿了一下，向大家问道：“在国内，数据挖掘应用的状况怎样？”

T钢铁公司的李部长抢先答道：“在我国，数据挖掘在互联网、金融、电信和商业等领域已经有一些成功的应用，而在其他行业如制造、航空、医药、反恐和刑侦等只有少量的尝试。”

“李部长的评价比较客观，但大家想过没有，为什么我们与发达国家的差距就这么大呢？”徐教授反问道。

教室里一阵沉默。

于是，徐教授坦率地表达了自己的看法：“其实我也一直在考虑这个问题，当然这里面的原因很多。直到你们管院张院长请我给你们上数据挖掘课时，我又发现了一个不可忽视的因素——政府和企业高层对数据挖掘不甚了解而导致他们对此不够重视或不能站在一定的高度提出有价值的需求。”

徐教授的一席话引起了李部长的共鸣，激动地说：“是的，徐教授讲得太对了。就拿我们钢铁公司来说吧，这几年，我们整天喊‘挺进世界500强’，忙于引进国外先进设备扩大生产规模，但却忽视与外界的技术交流而成为井底之蛙，就连数据挖掘这样在世界500强企业如雷贯耳的新技术我们却闻所未闻。由于自己不具备这方面的知识，生产管理中遇到了不能解决的问题，自然不会用数据挖掘的思想思考，甚至基层部门提出使用这样的方法，领导层却因对此不甚了解而不给力支持。”

李部长的话说到了其他学员的心坎上，他们个个首肯。

徐教授走下讲台，语重心长地说：“所以，我给你们上数据挖掘课来了，我期望从领导普及数据挖掘知识开始，唤起人们对数据的新认识，使你们告别基于简单统计分析的‘报表’决策时期，跨入使用数据挖掘技术的‘知识’决策时代。你们这些社会各界的精英们肩负的历史责任太大了，不管是政府部门的领导还是企业的老总，你们每天都在做各种各样的决策，稍有不慎就可能给国家和企业带来重大损失。我相信各位想为国家贡献自己的力量，但陷入‘心有余而力不足’的境地，正所谓‘我们沉浸在数据的海洋，渴望知识的淡水’！”



听完徐教授一席话，下面的各位老总感慨颇多，台下一片沉思。

徐教授鼓励大家道：“数据挖掘的最高境界就是‘**从数据中获取知识，辅助科学决策**’。希望通过我们的数据挖掘课程的学习，使你们了解到**什么是数据挖掘**？**它能够干什么**？**有哪些数据挖掘技术**？**怎么应用**？大家要认识到，数据挖掘不同于一般的管理软件，编好了拿来用就是了，数据挖掘在行业的成功应用也是一种创新。其实在数据挖掘算法方面，国内（也包括我）的研究团队也有一系列的国际水平的研究成果，但愿我们一起共同努力，推动数据挖掘技术在各行各业的应用，为建设创新型国家做出最大的贡献！”

教室里，又是一阵激动人心的掌声。

徐教授摆了摆手，接着说：“不过，给你们上这门课可让我费了不少脑筋，你们这些学员走向工作岗位都在10年以上了，大学所学的数学知识大都还给了老师，针对

研究生的讲法对你们不适用了。不过，我想出一种专门针对你们的案例教学法，通过典型的应用实例深入浅出地介绍数据挖掘的概念、功能、流程和算法。”

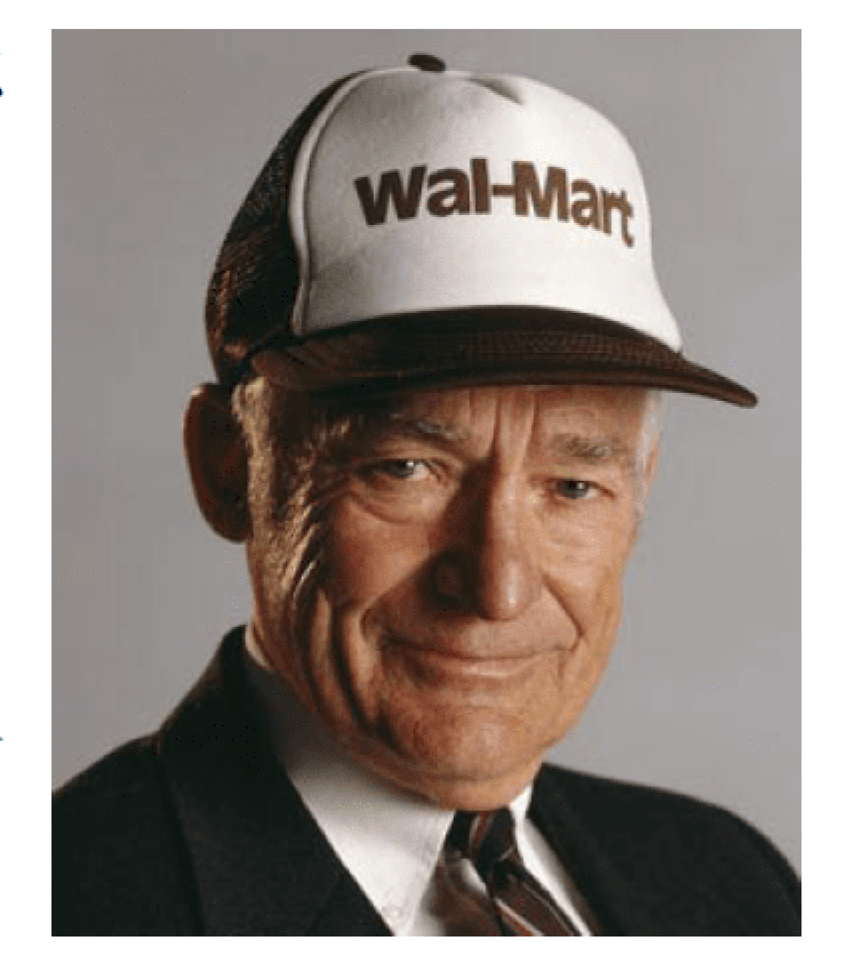
“太好了，徐老师。我曾经翻过几本数据挖掘的书籍，但理论性太强，满篇数学公式，真让人望而却步，而且应用实例甚少，让人难以理解。”李部长感慨地说。

徐教授接着说：“OK，言归正传，让我们开始数据挖掘之旅吧。我先给大家讲三个真实的故事，让你们感受一下数据挖掘到底是神马还是浮云？”

1.2 数据挖掘的故事

1.2.1 震撼业界的发现

“有一个人叫萨姆·沃尔顿的人，大家认识吧？”徐教授问道。



教室里鸦雀无声。

“那沃尔玛，谁没听说过？”徐教授接着问。

“连三岁小孩都知道。”一学员小声说。

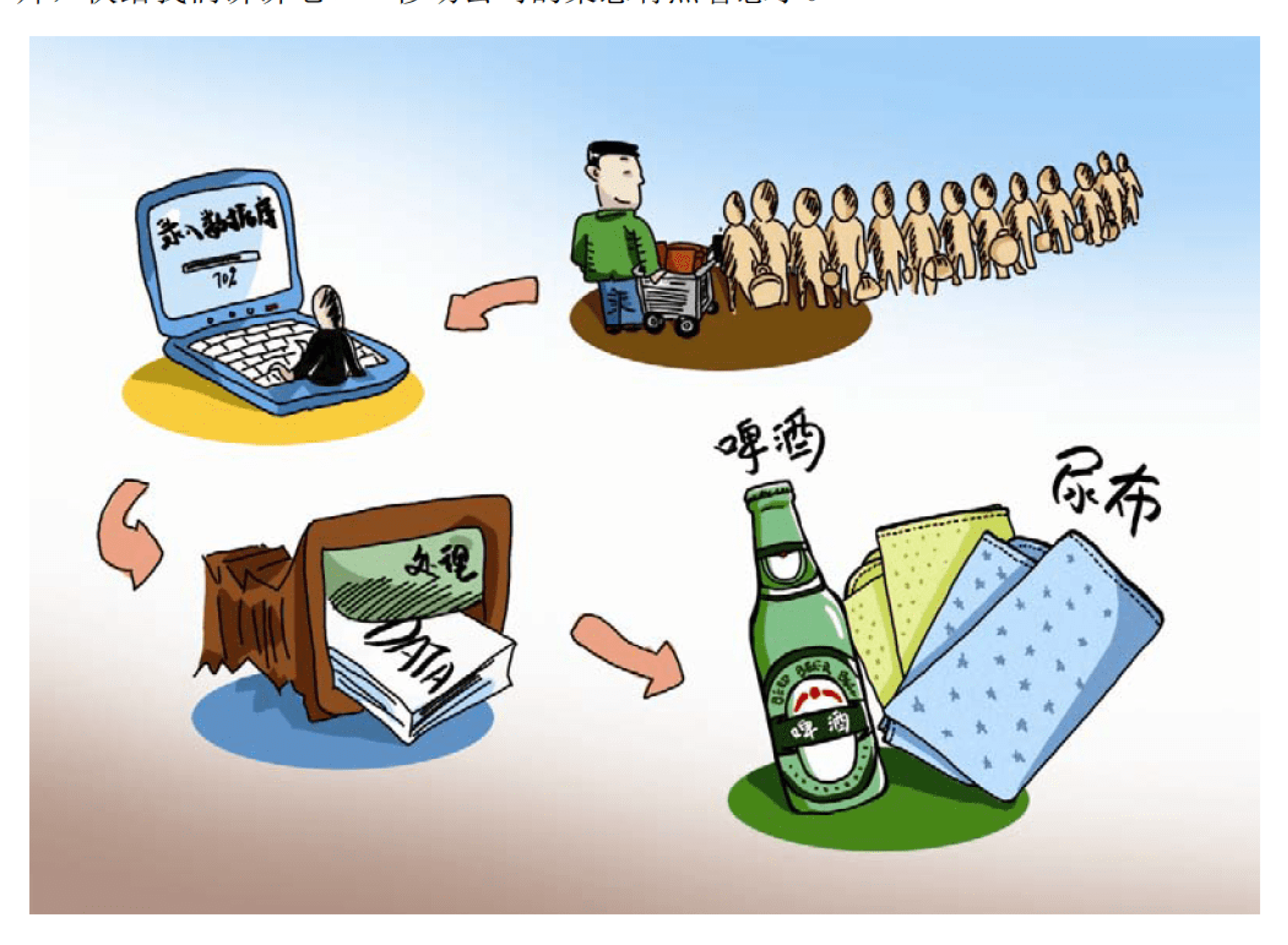
“哈哈，萨姆·沃尔顿是沃尔玛公司的创始人呀！”徐教授笑着说。

“对了，想起来了，萨姆·沃尔顿，是他将一个百货商店奇迹般地经营为全球最大的连锁零售企业，早在 1985 年 10 月就被《福布斯》杂志列为全美富豪排行榜的首位，连美国前总统布什都赞扬他是地道的美国人，展现了创业精神，是美国梦的缩影……”某超市的万总补充说。

“是的，勤奋、创新是这位智慧商人成功的法宝。他的‘日落原则’、‘十英尺

态度’和‘三米微笑’等服务理念以及营销策略‘女裤理论’和‘啤酒与尿布’至今在商业界令人津津乐道。更令人难忘的是，本世纪初**‘啤酒与尿布’**简直就成了‘数据挖掘’的代名词。”徐教授继续说。

“啤酒与尿布，这两个风马牛不相及的东西怎么与数据挖掘扯上了关系？徐老师，快给我们讲讲吧！”移动公司的梁总有点着急了。



“1983 年，当一般零售商还在进行信息化建设的时候，沃尔玛已经开始与休斯公司合作，花费 2400 万美元发射了一颗人造卫星，此后先后投入 6 亿多美元建起了电脑与卫星系统，还发明了条形码、无线扫描枪、计算机跟踪存货等新技术。借助于整套的高科技信息网络，沃尔玛的各部门沟通、各业务流程可迅速、准确地运行，数据库系统很快积累了海量的经营数据，包括大量的顾客消费行为记录。一年一度的圣诞节快要到了，沃尔玛人按照惯例又一次筹划节日的营销策略。这一次他们使用了一种新的‘购物篮分析’软件，对海量的顾客消费行为进行分析，一个意外地

发现让他们瞠目结舌，‘跟尿布一起购买最多的商品竟然是啤酒！’”

“这怎么可能呢？”有学员也感到疑惑不解。

“经过反复计算、核实，结论没有错。”徐教授答道。

“不过，这个故事告诉我们什么？”又有人问道。

“告诉我们数据挖掘可以发掘埋藏在海量数据中有价值的信息。”徐教授答道。

突然，后排有人大声说：“也告诉大家如果想喝啤酒，老婆不让买，就说去买尿布吧！”惹得大家哄堂大笑。

接着，徐教授问：“这是数据挖掘技术对历史数据进行分析得出的知识，这个结果符合现实情况吗？是否有利用价值？”

“还利用价值，真是六月里穿皮袄——反常！”有学员不以为然。

“紧接着，沃尔玛派出市场调查人员和分析师对这一结果进行了深入研究，证实它揭示了一条隐藏在‘尿布与啤酒’背后的美国人的一种行为模式：一些年龄在25～35岁的年轻父亲下班后经常要到超市去给婴儿买尿布，而他们中有30%～40%的人会顺手为自己买几瓶啤酒。”

刚才那位学员想通了，小声说：“对了，这是在美国，老外的行为模式与中国人就是不一样！证实了这样的发现是符合实际的，沃尔玛会怎么办呢？”

徐教授挥动了一下电子教鞭，大声说：“沃尔玛立即采取了行动，将卖场内原来相隔很远的妇婴用品区与酒类饮料区的空间距离拉近，使顾客更加方便。然后对本地区新生育家庭的消费能力进行了调查，对这两个产品的价格也做了调整，并向一次购买达到一定金额的顾客赠送婴儿奶嘴及其他小礼品，结果是尿布与啤酒的销售量双双大增。”

某超市的万总激动地站了起来，情不自禁地说：“不愧为全球零售业巨头啊，高招，值得借鉴！”

徐教授一边示意她坐下，一边说："是的，不仅在零售业值得借鉴，这种'购物篮分析'后来演变为'关联规则分析'，并在其他行业发挥重大应用，我们 EMBA 班的学员有很多来自于工业界，下面再给你们讲一个工业生产中利用数据挖掘技术节约成本的故事。"

1.2.2　降低成本的绝活

徐教授："工业界的学员都知道，派克汉尼汾公司是一家世界一流的工业企业，总部位于美国，于 1918 年由 Arthur L.Parker 先生创立。早在上世纪 70 年代已发展为全球控制领域最广、产品种类最完备的公司，年销售额超过 100 亿美元。大家估计下派克公司的年维修费用是多少？"



"200 万美元？"

“500 万美元？”

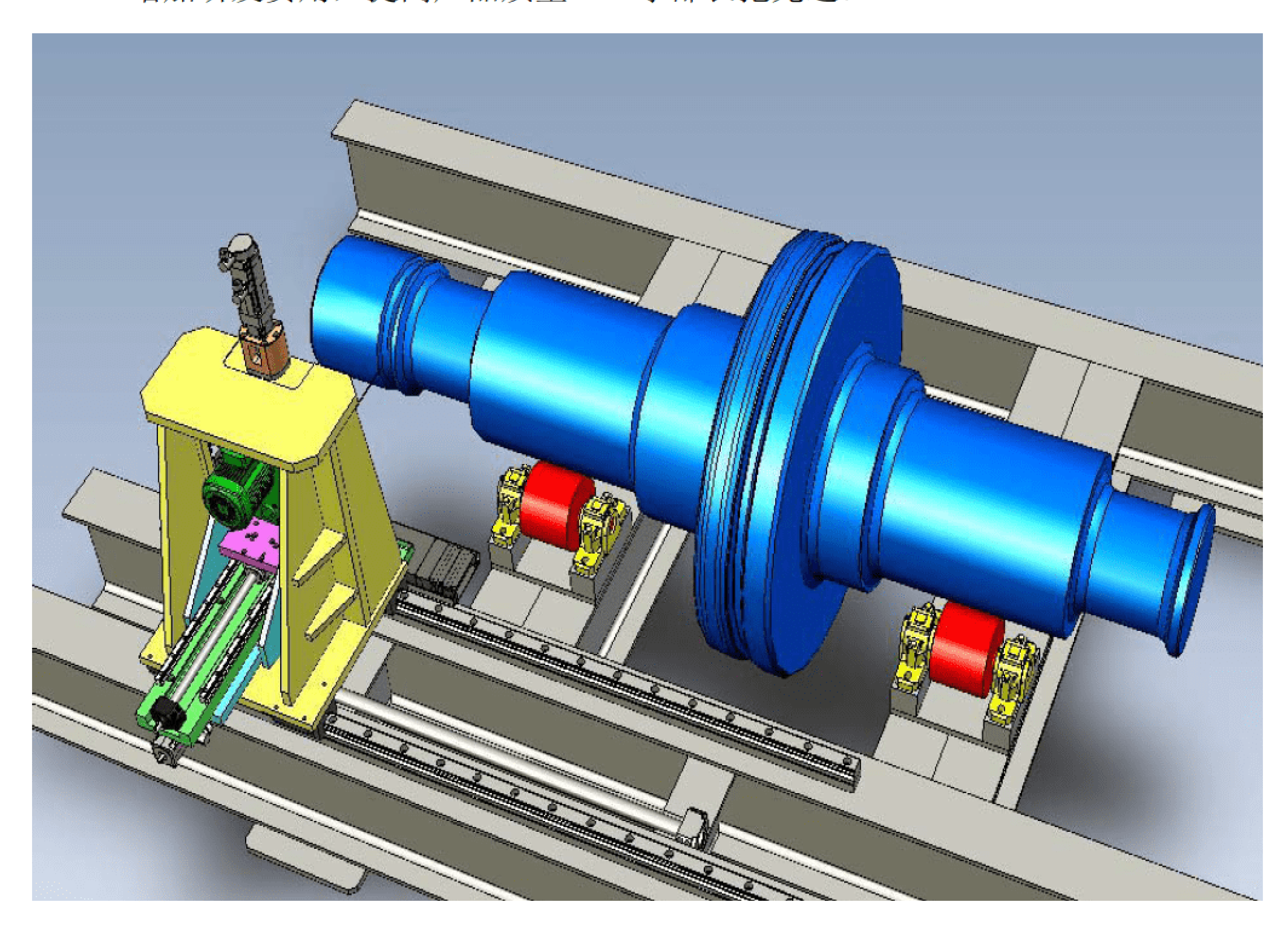
……

“该公司产品出售后保修一年，年维修费用超过了一亿美元。” 徐老师说。

“我们鼓风机厂的年产值也比不上人家的年维修费。”一学员喃喃自语。

“那怎样降低维修费用呢？”徐教授问道。

“增加研发费用，提高产品质量！”李部长抢先道。



“不错！但是如果我们假设在目前的技术条件下，产品质量已经达到了较高标准。还有没有其他办法？”

“这个……难道是数据挖掘？”有一学员自语道，其他学员则低头沉思。

徐教授肯定地说："是的，派克公司采用了数据挖掘方法。以一款干燥器为例，该机器 1200 多种零件中，常坏的贵重零件约 20 种。应用数据挖掘的关联规则分析发现这些价格昂贵的零件的寿命竟然大多数与少数几种便宜零件的磨损有关。"

李部长激动了："妙，妙极了。采用常更换便宜部件，达到延长贵重部件的使用寿命，就可以大大地降低维修成本。我们怎么就想不到呢！"

徐教授看着李部长，说道："对了，派克公司采用了这样的策略后，仅在这干燥器这种产品上，每年节省维修费高达上千万美元。"

李部长坐不住了，大声说："我们公司的不锈钢生产线也有同样的问题。徐老师，您指导我们也挖掘挖掘吧！"

徐教授："别着急，李部长。有很多数据挖掘方法能够解决你们公司生产管理、新产品设计、产品质量控制、能源分析、原料搭配、成本分析等许多问题，以后我们再进一步讨论。"

大家越来越坚信数据挖掘的巨大威力，精神也更加集中了。

1.2.3　出奇制胜的小纸条

徐老师接着说道："我们在座的学员大部分喜欢看足球比赛，我再给大家讲个数据挖掘在体育方面应用的故事。"

这时，PPT 上出现了一个章鱼，光笔的红点在它身上晃动，徐教授问道："上届世界杯的时候名噪一时的'章鱼帝'大家还记得吧？"

"出道两年的章鱼保罗在 2008 欧洲杯和 2010 世界杯两届大赛中，预测 14 次猜对 13 次、成功率 92%，堪称不折不扣的'章鱼帝'。"足球迷李部长先吐为快。

徐老师补充道："从科学的角度来看，章鱼帝的预测仅是小概率事件在万众瞩目下发生了而已。但是 2006 世界杯同样是德国和阿根廷的赛场上，不是章鱼保罗救了

德国，而是一个神秘的小纸条。”

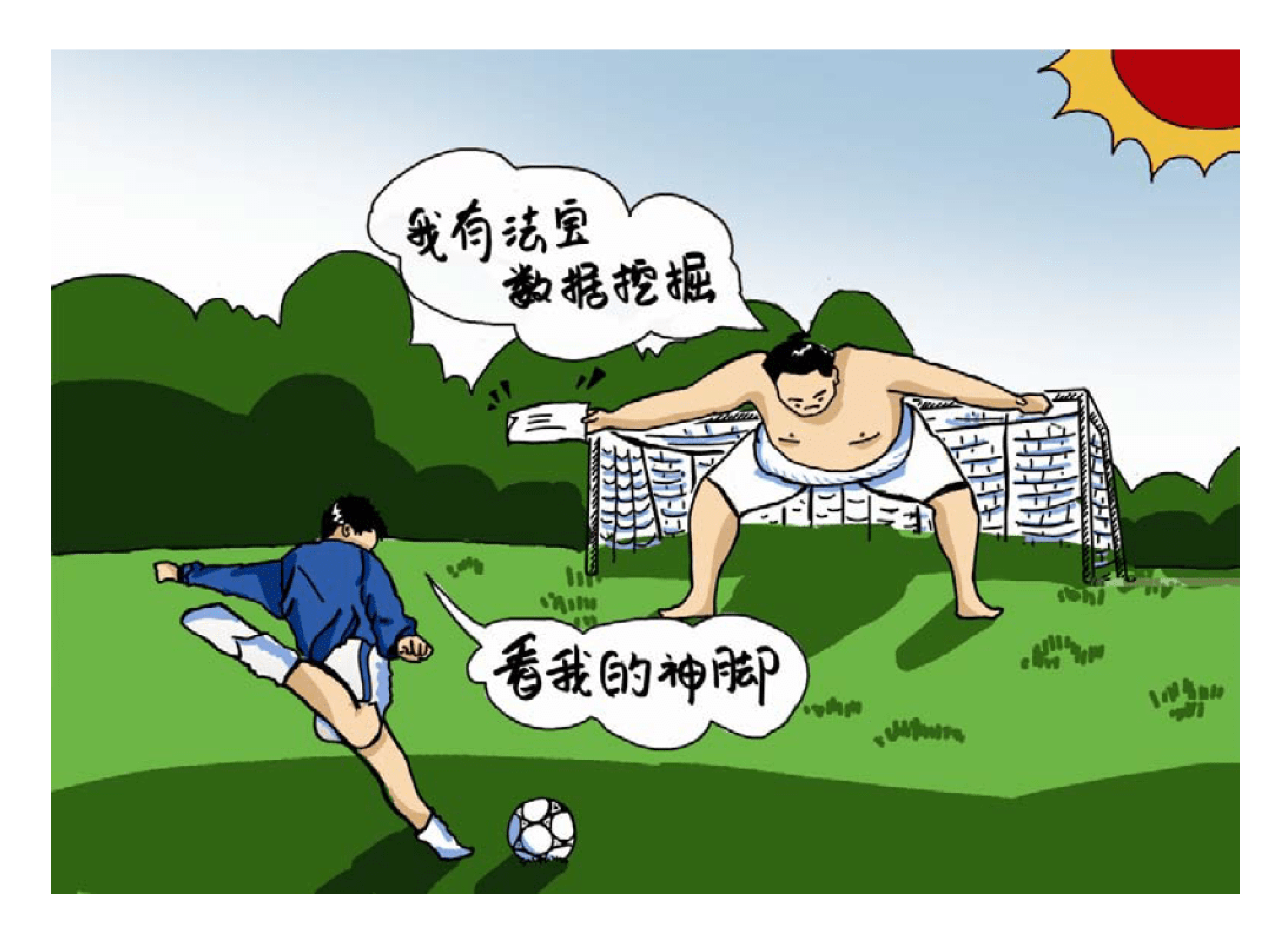
“一个小纸条有这么大的作用，到底是什么小纸条啊？徐老师您赶紧给我们讲讲吧！”有人急不可待。



徐教授不紧不慢地说：“2006 年世界杯上，阿根廷和德国在 1/4 决赛中 120 分钟难分高下，在点球大战之前，老门将卡恩将一张纸条递到莱曼手中。莱曼每次扑点球之前都要看一眼纸条。结果是，莱曼所有点球都判断对了方向，除了两个点球质量太高无力回天外，其他全部扑出，阿根廷只能黯然出局。”

“那纸条上到底写着什么锦囊妙计？”

“写着德国胜！哈哈，可惜章鱼保罗还没出生。”台下哄笑一堂。



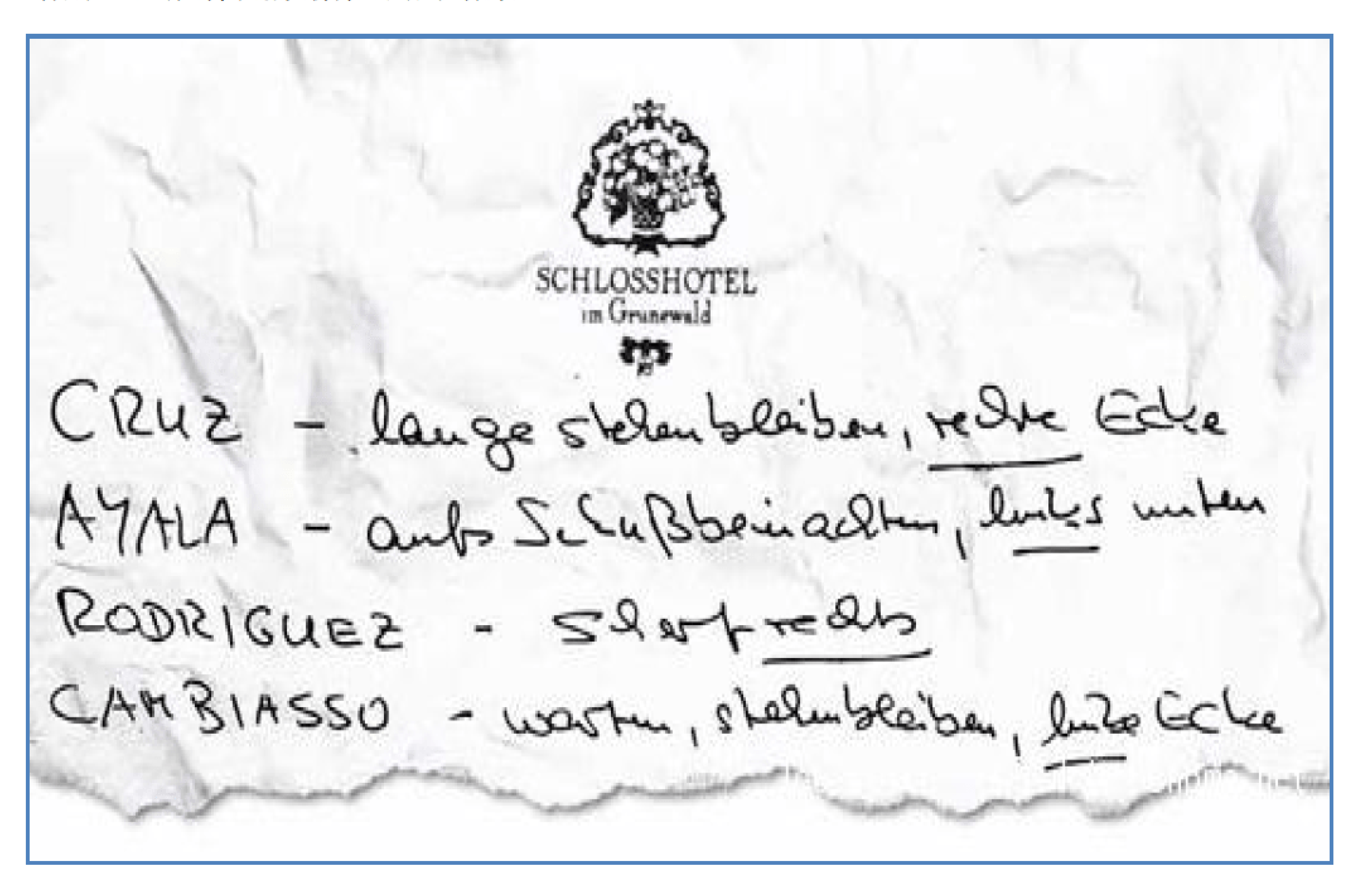
徐教授摆了手势，让大家安静，继续说道："上面记录着阿根廷队的克鲁兹、阿亚拉、罗德里格斯以及坎比亚索习惯的脚法。德国队守门员教练科普克如此精确地预测出阿根廷球员射出的点球方向，并不是他有什么过人的占卜天才。那张草草写在格鲁内瓦尔德皇宫酒店便笺上的扑点球秘籍，来自于德国科隆体育学院数据分析小组夜以继日的努力。"

"点球就是点球了，纯技术问题，有什么可分析的嘛？"足球迷李部长不以为然。

徐教授："这个问题问得好。分析小组的人员收集了阿根廷队 13000 个点球的录像，所有这些采集回来的点球数据被输入数据库中，并根据阿根廷射门练习的数据找出了一些可以描述射门动作的行为特征，最终从这些特征中提炼出很少的更具体特征。大家说说点球动作行为特征可以分为几类？"

"两类，进球和没进球！"某人的幽默引来全班大笑。

徐教授补充道："这些特征被描述为：阿亚拉，短助跑，右下角；里克尔梅，斜向助跑，右下角；马克西，长距离助跑，左上角；坎比亚索，长距离助跑，右侧；索林，短助跑，右下角；特维斯，短助跑，中路……。这些特征描述了阿根廷队谁罚点球、怎样罚点球的规律。正是这张小纸条把大力神杯交到了德国队手中！小纸条上总结的这些规律是数据挖掘的结果！"



某省鼓风动力集团的王总快人快语："数据挖掘可太有用了。徐老师，您快给我们讲讲什么是数据挖掘吧。"

这时，下课铃响了，徐教授示意大家休息。

1.3　什么是数据挖掘?

新的一节课开始了，徐教授走上了讲台，清了清嗓子，声音更加洪亮："随着

计算机技术、数据库技术、传感器技术和自动化技术的飞速发展，人们获取数据、存储数据变得越来越容易。这些数据不是人为产生的，是对我们所研究对象隐含的一定规律的反映。数据挖掘的目的就是要从所获取的数据中发现这种规律性的知识，从而帮助企业在他们的数据仓库中找到最重要的信息，预测未来趋势和行为，使得商务和生产活动具有前瞻性，并作出具有知识驱动的决策。”

徐教授将 PPT 翻回到数据挖掘的故事，继续说：“通过上节课所讲的三个故事，相信在座的同学对数据挖掘有了初步的认识。那么到底什么是数据挖掘呢？大家可以发表下自己的观点。”

学员们你一言，我一语，争先恐后。

“数据挖掘就是从数据中发现有价值的信息的技术。”

“数据挖掘是对数据建立模型，通过算法求解而发现隐藏在数据中的知识的一种手段。”

“……”

徐教授总结道：“大家对数据挖掘的认识都值得表扬，不过表述得都不够全面。”说着，徐教授敲了一下键盘，说：“请看大屏幕，这才是最权威的数据挖掘的定义。”

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用信息和知识的过程。

大家认真地看着屏幕的内容。

片刻之后，有学员问道：“数据量小是不是就不能进行数据挖掘了？”

徐教授答道：“实际上数据挖掘的算法大都是建立在统计学**大数定律**基础上的。数据量太小，常常无法反映出真实世界中的普遍特性，这样挖掘算法得出的结论自然不可靠。但并非小数据量就不可以进行挖掘，近年来研究者也提出了一些对小样

本进行挖掘的方法，如支撑向量机方法就是基于小样本学习理论的非常实用的方法。数据量虽小，但数据总是事物特性一定程度的反映，只要建立的模型和算法得当，当然也可以从这些数据中获取一定的信息。”

“那么是不是数据量越大越好？”有学员问。

“从理论上说，应该是这样。但随着数据量的增大，算法执行效率会越来越低，甚至无法计算。”徐教授回答说。

刚才提问的学员点了点头，接着问：“徐老师，数据挖掘的定义中，数据前面还有那么多的修饰，您还是给我们解释解释吧。”

“大家淡定点，‘不完全的、有噪声的、模糊的、随机的’确实有点绕口，现实中经常会碰到这种数据。例如，问卷调查时发现不少人不填婚姻状况和年龄，这些**不完全的或缺失的**数据会给数据挖掘带来一定的难度，我们要么干脆删除这些样本或记录，要么选择使用一定的方法将这些缺失数据补上，或者选择使用可以自动处理缺失数据的算法。”说到这儿，徐教授端起了茶杯，说自己也要补充一下水分了。

“那噪声是什么意思？”一个学员问。

徐教授合上茶杯盖子，一边狠狠地用杯子连续敲击着桌子，一边说：“对于我讲课的声音来说，敲桌子的声音就是噪音，我们的录音机录到的是我的讲话声和敲桌子声混杂在一起的混合声波数据。”

“我明白了，由于异常情况的干扰，使我们获得的数据偏离了真实值，这样的数据就是噪音数据。”刚才提问的学员说。

“不光是外界的干扰，测量仪器的故障、人工输入或抄写时的失误等都可能形成**噪音数据**，可见实际问题中噪音数据往往难以避免。”徐教授进一步解释说。

“徐老师，什么是模糊的、随机的数据？”又有一学员问。

“在数据挖掘过程中，我们不可避免地要涉及事物的不确定性。不确定性包括

模糊性和随机性。**模糊性**则指事物本身从属概念的不确定性，**随机性**是指事件发生与否的不确定性。”

“太抽象了，徐老师，您给我们举个例子吧！”李部长建议说。

“好吧。其实模糊的数据大家平时经常见到，比如说张三个子很高，李四个子较矮，个子的高矮就是典型的模糊性概念，到底多高才算高，李部长 1 米 80，对一般人来说算高个子，但跟姚明比，就太矮了。随机数据也极为多见，比如说超市啤酒每天的销量显然是不确定的，大部分人买啤酒是在超市转悠时临时决定的。”徐教授回答道。

李部长扶了扶眼镜，支支吾吾地说：“我似乎明白了……”

本科应用数学专业毕业的王总快人快语：“李部长，我借给你《模糊集的应用》和《概率统计》两本书，看看你才会真正明白。我要问新的问题了, 徐老师, 数据挖掘的目的是从数据中发现新的信息和知识，那挖掘出来的知识是什么？”

徐教授回答道：“挖掘出来的知识就是‘散落的珍珠’，亦或是‘发光的金子’，它的实际决策价值非凡。知识是通过对数据进行深入地归纳、分析而获得的，是对所研究对象更深层次的认识。知识是隐藏在数据中的关于所研究对象的一种规律性，比如可以用来预测的数学模型、‘如果……那么……’这样的规则、描述事物的类别、有价值的模式、所研究对象的结构、研究对象与对象之间的关系等。”

1.4　历史的必然

EMBA 教室的座位是半弧形的，中间有通道，老师讲课时部分时间是站在学生中间的，课堂上师生交流非常方便。

“人类走过了石器时代，纸器时代，磁器时代，直至现在的网络技术时代和正在跨入的物联网时代，这些智慧、文明的结晶是怎么样代代相传，生生不息地保留和继承下来的呢？”徐教授问。

“信息获取……”

“信息存储……”

“信息查询……”

“信息的加工和应用……”

旁边的学员们陆陆续续地表达了自己的看法。

“对，确实是这样。人们通过信息的获取、存储与查询、加工和应用几个环节实现知识传播、继承和发展。”徐教授对学员们的回答很满意。

随后，徐教授通过 PPT 展示了一个图，并讲述了伴随着人类历史文明发展和进化的长河，人们对知识和信息的存储、加工应用的演化进程。



“从人类有了获取信息的能力开始，便不断对信息进行归纳总结。大家想想，有

哪些谚语可以说明，古人就开始针对观察到的信息进行分析和归纳了？”徐教授刚问完，谚语大接龙便开始了。

“连发三日东北风，定有大水后面跟。”

“天上起了泡头云，不过三天雨淋淋。”

“星光闪闪如动摇，大雨下得没处逃。”

“……”

课堂气氛一下子变得十分活跃，大家在说起古人智慧的时候，都觉得万分光荣和自豪。

“通过祖祖辈辈的观察、积累与归纳，人们发现了自然现象与天气的‘关联规则’”，徐教授总结说。

突然，第一排的一个学员站起来说道：“对于一些简单的自然现象，可以通过归纳提取形成经验知识，但现实世界太多的复杂问题，数据量极大，已经远远超出了人脑可处理的范围。”

他旁边的一位学员也感慨地说：“是的，现在获取数据非常容易，就拿我们钢铁公司来说，每日产生的数据超过3Gb，要是将这些数据放在我的脑子里，脑瓜肯定爆炸了，更不用说处理、归纳得到知识了。”

看看他憨憨的笑容，大家都被逗乐了，之后便都陷入了沉思。

“不是有计算机么，人就不用操那么多心了。”另外一个学员小声说。

于是，徐教授解释说：“上世纪60年代，尽管有了计算机，但对数据是以零散文件方式进行管理的。我们能够收集、存储、处理如此海量的数据，归功于20世纪70年代IBM发明的关系式数据库和SQL查询语言。在此基础上通过计算机和网络进行联机事务处理（OnLine　Transaction Processing，OLTP）可以对管理信息进行日常操作并及时、安全、高效地存储数据，这样便引发了数据爆炸式地增长。”

电信公司冯总，计算机专业硕士，在单位负责数据仓库建设，听到这里，话匣子关不住了："OLTP关心的只是业务操作，只对当前数据感兴趣。其实信息处理的目的是为人们提供决策支持，这就需要对历史数据进行大量地分析处理。对历史数据的分析，往往导致系统进行长时间运行，严重影响日常数据实时操作，这就要求把分析性操作及其相关数据从事务处理环境中提取出来，按照决策支持的需要进行重新组织，建立单独的分析环境。"

李部长这几年读了不少信息处理方面的书籍，他接上了话茬："为了满足这种需求，W. H. Inmon 于1993年出版了'Building the Data Warehouse'，从此数据仓库（Data Warehouse）隆重登场，W.H.Inmon也当之无愧地成为数据仓库之父。他给出了数据仓库定义：**'数据仓库是一个面向主题的、集成的、随时间变化的、持久的数据集合，用于支持管理层的决策过程'**。在数据仓库产生的同时，**联机在线分析（OnLine Analytical Processing，OLAP）出现了，它是一种具有对数据进行汇集、合并和聚集以及从不同角度观察信息的分析技术。**"

电信公司冯总，在单位里被誉为数据仓库专家，继续说："通过OLAP技术可以对从数据库或数据仓库得到的经验、规则进行验证，当然也可以对数据挖掘结果的有效性、可行度进行检验、完善。然而，数据库和数据仓库越建越大，通过直观的感觉、简单的统计分析和OLAP技术并不能发现隐藏在数据中有价值的信息和知识。"

"上世纪80年代末到90年代初，广泛流传着一句耐人寻味的话'我们沉浸在数据的海洋中，但却渴望着知识的淡水'，这句话生动地描绘了人们面对海量数据的迷惘和无奈。"徐教授深沉地说。

突然，徐教授抬高了嗓门："一石激起千层浪，这时沃尔玛演绎了一场'啤酒和尿布的故事'，它使人们看到了数据分析的希望，擂起了数据挖掘的战鼓, 一场数据挖掘的风暴开始了……"

几个学员抑制不住内心的激动，你一言、我一语地表达自己的观点：

"商业界发现了沃尔玛迅猛发展的密招，纷纷效仿。"

“电信行业沸腾了，各公司纷纷争先恐后地利用数据挖掘这一锐利武器解决他们面临的最紧迫的问题，如客户分群、客户流失原因及预测、业务套餐及响应、关联消费等。”

“工业界也着急了，他们的数据堆积如山，期望从中挖掘出金子，指导生产和管理。”

“科学界大批科研工作者聚焦于数据挖掘，紧锣密鼓地投入到该新生领域的研究。”

“……”

徐教授走上讲台，总结道：“人常说，‘需求’是成功之源。商业管理、生产控制、市场分析到工程设计、科学探索等将堆积如山的数据资源转换为信息和知识的巨大需求，促使着数据挖掘技术的飞速发展。九十年代中期以后，基于数理统计、人工智能、机器学习、神经网络等多种技术，关于数据挖掘软件的开发和应用成为热点。”

徐教授的话音刚落，有学员便问道：“徐老师，您一会儿说数据库中的知识发现，一会儿又用数据挖掘，我真不知道它们之间的关系。”

“2008 年我在李部长他们钢铁公司作数据挖掘报告，也有几个人问我同样的问题。在 1989 年 8 月第 11 届国际人工智能联合会议上，数据挖掘以数据库中的知识发现（Knowledge Discovery in Database，KDD）第一次正式亮相。从此以后，**数据挖掘（Data Mining）和数据库中的知识发现（KDD）互为别名，但后来数据挖掘渐渐被多数人喜闻乐道**。”徐教授回答道。

刚才那位提问的学员是 EMBA 班里有名的“问到底”，继续穷追不舍：“徐老师，数据挖掘是在什么时候被大家普遍接受的呢？”

李部长急了，站了起来：“‘问到底’同学，你是不是一定要考倒徐老师！”

徐教授赶紧解围：“这个问题已经难以考证，大约在上世纪 90 年代开始，数据挖掘占了上风，其中还有一段趣事。”

“问到底”顾不上理会李部长，高兴地说：“徐老师又要讲故事了。”

徐教授示意李部长坐下，笑着说：“其实学院派最初一直沿用数据库中的知识发现即 KDD。在一次 KDD 国际会议中，委员会曾经展开讨论，到底使用 KDD，还是 Data Mining。”

“问到底”急切地说：“肯定一致同意使用 Data Mining。”

徐教授摆了摆手道：“会议上大家争论不休，讨论了两个小时没有结果。要是你们是当时参会的专家，会怎么定这个名字？”

“抓阄”

“抛硬币”

“听会议主席的”

“……”

学员们也开起了玩笑。

徐教授说到：“呵呵，我们中国人喜欢举手表决，外国人也兴这一套。会议主席最后决定投票表决，结果很具有戏剧性，一共 16 名委员，其中 8 位投票赞成 KDD，另 8 位赞成 Data Mining。”

“问到底”露出一副为难的表情：“这可怎么办呢？”

徐教授答道：“事实上，根据当时会议的记录，最后一位元老站出来说‘数据挖掘这个术语太为土气，科学研究就是要获得新的知识’。”

“问到底”感到有些失望：“老奸巨猾，跟没说一样！”

“怎么跟没说一样？他作了双重肯定。于是在**科研界便继续沿用 KDD 这个术语，而在商用领域**，因为‘数据库中的知识发现’显得过于冗长，**就普遍采用了更加通俗、简单的术语‘数据挖掘’**。”

1.5　数据挖掘能干什么?

要讲数据挖掘的功能，大家都非常感兴趣。徐教授提高了嗓门：“前面我给大家讲了数据挖掘的三个故事，并给出了数据挖掘的定义，还简要地回顾了一下数据挖掘产生的过程，可数据挖掘到底能干些什么呢？”

“购物篮分析”

“用户分群”

“客户流失分析”

“服务套餐设计”

“预测”

“……”

学员们纷纷根据自己的直观理解回答着。

“大家所说的只是根据我前面讲的内容概括了数据挖掘的一些功能。有个成语叫做‘盲人摸象’，我才领着大家摸了大象的一条腿而已，哈哈。”徐教授开玩笑。

“徐老师，在座的学员大部分是政府部门和大中型企业的头头脑脑，我们首先希望知道数据挖掘到底能够干什么，至于怎么干那就是工程师的事了。您就先概括一下数据挖掘的功能吧。”高新区的段主任建议说。



徐老师：“好吧。概括地说，数据挖掘的功能主要包括关联分析、聚类分析、分类、回归、时间序列分析和偏差甄别等，下面我们分别介绍这些功能。”

1.5.1　关联（association）规则挖掘

徐教授又将 PPT 翻回到“啤酒与尿布”的画面，说：“大家还记得吧，沃尔玛在海量的交易数据中发现了美国人的一种行为模式：年龄在 25～35 岁的年轻父亲在给婴儿买尿布的同时，有 30%～40%的会为自己买啤酒。这就是轰动一时的啤酒与尿布的关联规则。”

听了徐教授的这句话，李部长灵机一动：“徐老师，这么说之前您讲的第二个故事中，派克汉尼汾公司发现昂贵零件的寿命与少数几种便宜零件的磨损有关也是一种关联规则。”

“对，关联是指一个事件与另一个事件之间的依赖关系。关联规则挖掘就是发掘数据库中的关联关系，大家还了解到哪些关联规则的应用？”徐教授问。

华润超市市场营销主管万总抢先说道：“徐老师，据我所知，关联规则已经成为各大超市安排商品布局、促进销售量的一种法宝。近年来，电信公司、保险公司和美容公司等服务行业都争先恐后地效仿零售业的这种做法，纷纷设计各种套餐，实现捆绑促销。”

电力公司的赵总：“在电力行业，一些发达国家通过关联分析对输变电设备进行状态检测，为状态检修计划的制定提供科学依据。”

卫生局江副局长：“国内外均有报道，有人将关联规则挖掘应用于临床疾病诊断，比如通过实例试图发现吸烟、环境污染、职业、肺部慢性疾病等因素与肺癌的发生之间的关联，从而发现肺癌与它产生的可能因素间的规则，利用规则模式指导肺癌的诊断与预防。”

“……”

大家纷纷介绍本行业中关联规则的应用情况，令徐教授惊诧不已，不解地问：“你们怎么都知道这么多？”

学员们含笑不语。

李部长道出了其中的奥秘：“徐老师，在 X 大学，谁都知道，您上课的最大特点是激情豪迈，互动共鸣。我们 EMBA 班的学员都工作了数年，现在能坐在教室充电，倍感机会来之不易，大家在您上课的头一天晚上都会进行预习并准备与您配合的材料。”

徐教授高兴地笑了，接着说：“那我就要再问了，最基本的关联规则挖掘算法是什么？该算法的基本思想是什么？”

教室里鸦雀无语。

徐教授环视了一周，发现华润超市的万总跃跃欲试，便鼓励说：“万总，你来说

说，不完全的我来补充。”

万总鼓足了勇气，大声道：“**最经典的关联规则算法是**由 Agrawal 和 Verkamo 于 1994 年提出的 Apriori 算法，此后近十多年来，这方面的文章已达上万篇之多，但都是基于这种算法围绕着如何提高关联规则挖掘算法的效率、在海量数据集上进行关联规则提取、如何挖掘有价值的关联规则和关联规则的应用这些主题进行研究的。至于 Apriori 算法的思想……，我记不太清楚了。”

徐教授鼓励说：“回答的不错，可见课前准备花了很大功夫，值得表扬。”

徐教授的话音刚落，万总又开了口：“我记起来了，Apriori 算法的基本思想是：**首先从事件中集中寻找所有频繁出现的事件子集，然后在这些频繁事件子集中发现可信度较高的规则。**”

徐教授示意万总坐下，继续说：“Apriori 算法的大概思想就是这样，算法的详细描述大家可参考教材。我想大家更关注的是关联规则的应用，近年来有很多学者开展关联规则与分类、聚类挖掘方法的结合研究；利用关联规则进行属性选择和数据降维等。我收集了一些这方面的研究成果和应用案例，请大家从我的个人网站下载阅读。”

1.5.2 聚类

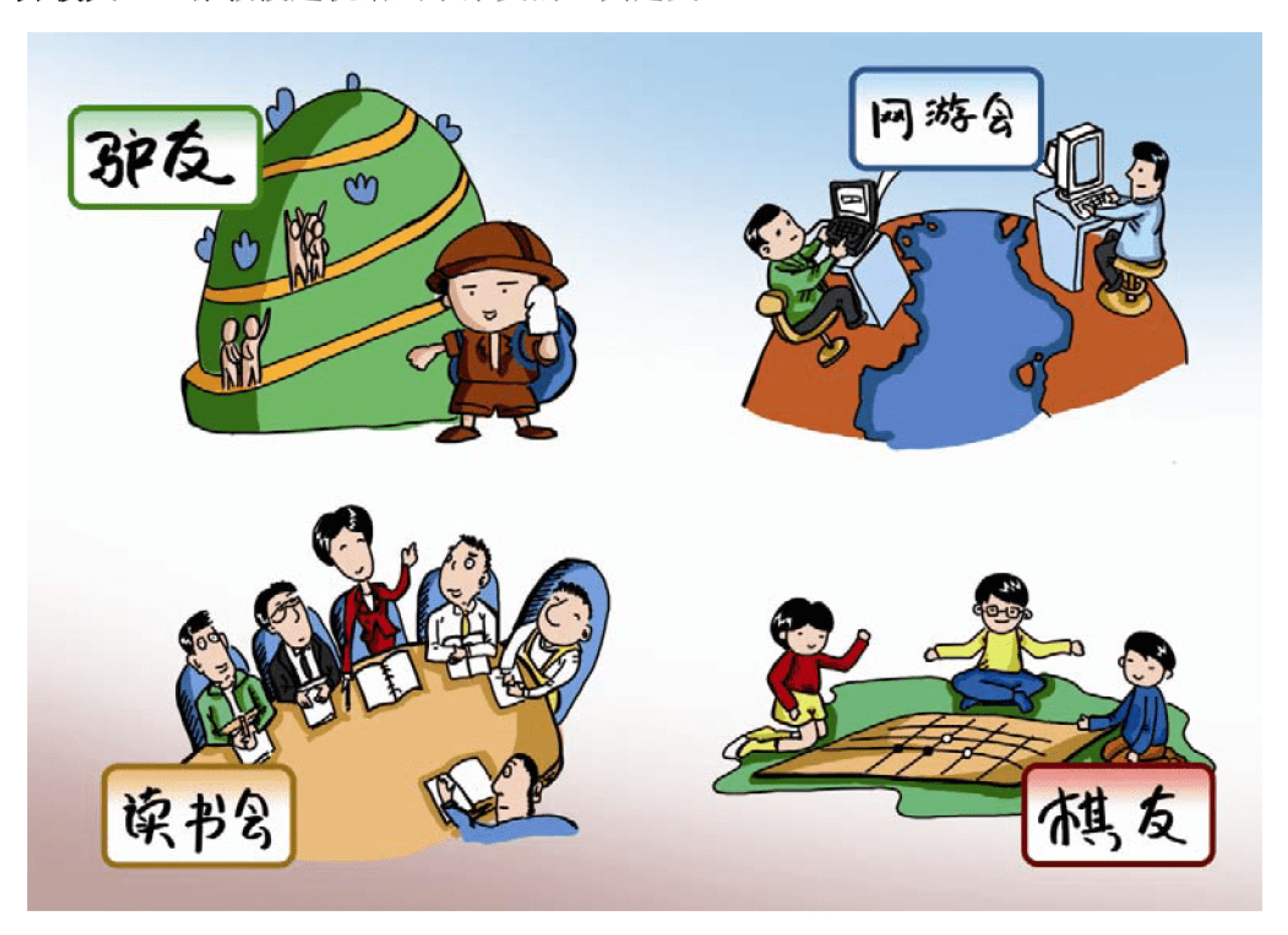
“在平时的人际交往和私下的生活空间中，大多数人会自觉不自觉地加入到一个个社交圈子中。‘驴友’、‘同学会’、‘高尔夫俱乐部’等，林林总总。真可谓‘物以类聚，人以群分’。”徐教授开始了聚类的讲解。

“徐老师，是不是说，圈子就是聚类？”一个学员问。

徐教授没有正面回答，继续说：“大家想一想，生活中的圈子有什么特点？”

李部长站了起来：“社会学家指出，‘圈子’就是由志向、趣味、地位、年龄、职业、爱好、特长、个性、收入甚至居住地点比较相近的人自发形成的团体。”

“对了，正是因为这些人具有相似的特征，他们才能聚集在一起。**聚类就是将数据对象划分成若干个类，在同一类中的对象具有较高的相似度，而不同类中的对象差异较大。**”徐教授趁机给出了聚类的经典定义。



刚才提问的那位学员从徐教授话语中悟出了聚类的真谛，感慨道：“我有点明白了，我们加入某个‘圈子’，实际上就是聚类的过程，因为这个圈子的成员与我们有着相似的特点。”

这时，徐教授才对这位学员的理解（圈子就是聚类）作了正面回应：“回答正确，加十分！”

“徐老师，从聚类的定义来看，进行聚类前并不知道所研究的对象有多少个类，聚类的过程就是通过相似性的度量，使对象聚集成若干个类，各个类的成员具有其共同的或相似的特性。”李部长说出了自己对聚类的理解。

徐教授认为李部长的理解已经比较深刻，频频点头。他因势利导，又提出了一个深刻的问题："聚类的关键是对象相似性的度量，大家想一想，如何度量数据对象的相似性呢？"

李部长抢答道："**两个对象间的距离越小，说明二者越相似，用距离度量对象的相似性应该是最自然的方法**。"

徐教授满意地点了点头："对，基于距离度量对象的相似性的思想，研究者提出了两类经典的聚类算法：**划分方法**和**层次聚类方法**。"

马处长似乎对这两种方法有所了解，说道："听我们数据挖掘算法组的小彭经常说 Partitioning Method 和 Hierarchial Method，原来就是指的这两类聚类算法。徐老师，昨天晚上我预习时大概了解了一下聚类算法，但理解不够深刻，您就给我们讲讲吧。"

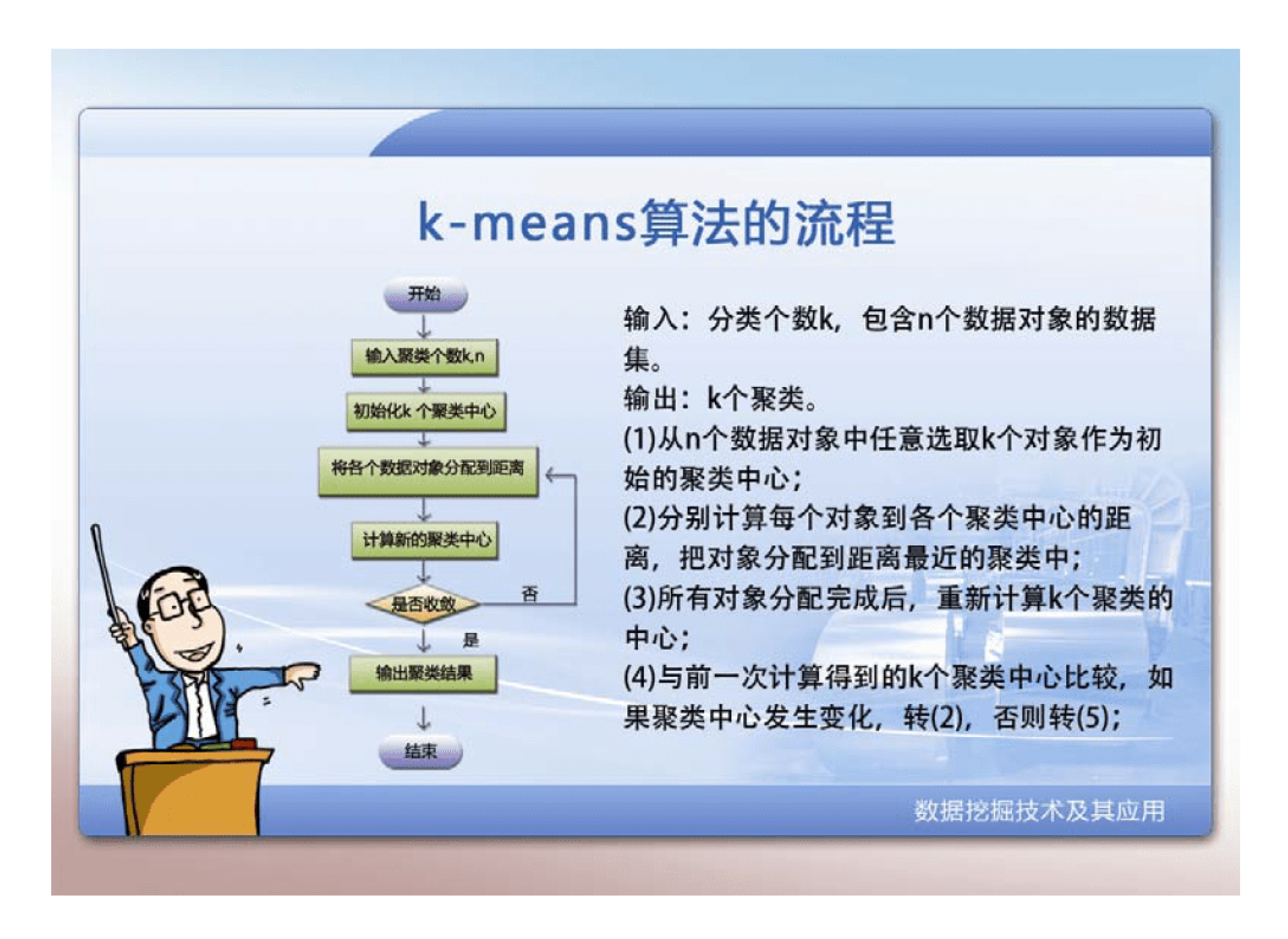
徐教授欣然答应，但没有立即开始讲算法，他先引导学员回顾基本的数学知识，问道："大家还记得距离怎么计算？"

电力公司的马处长简洁地答道："用欧氏距离呗！"

"对，就是大家在高等数学中经常用到的欧几里德（Euclid）距离。不过在聚类分析中，还经常用到曼哈坦（Manhattan）距离、切比雪夫（Chebyshev）距离、马哈拉诺比斯（Mahalanobis）距离等。其实，凡是满足距离定义的四个条件（即唯一性、非负性、对称性和三角不等式）的函数都可以作为距离公式。"

徐教授扫视了一下学员，觉得大家理解了距离的含义，于是说："好了，我现在就简单地介绍一下基于距离的聚类算法：划分方法和层次聚类方法。这两类方法的典型代表分别为 k-Means、k-Medoids 和聚集、分裂算法。下面我就分别介绍这些算法。"

徐教授翻动了一下 PPT，接着说道："k-Means 算法的核心思想是把 n 个数据对象划分为 k 个类，使每个类中的数据点到该类中心的距离平方和最小。"



李部长的脑子是双核的，徐教授的话音刚落，他便道出了他的理解："徐老师，k-Means 算法本质上是在实现**聚类的基本思想：类内数据点越近越好，类间点越远越好的尽可能算法**。"

"李部长理解得完全正确，不过 k-Means 算法的思想只是给出了一个优化目标——距离之和最小，具体实现一般使用如 PPT 图示的迭代算法。"

学员们都将注意力集中在 k-Means 算法框图上，马处长看出了问题："徐老师，k-Means 算法事先就给定了聚类的个数 k，然后通过迭代过程将数据点聚集到 k 个类中去。但是，一般情况我们并不知道数据点可以聚集成多少个类。"

"马处长说得对，k-Means 算法就是要尝试找出使平方误差函数值最小的 k 个划分，为了找出最合适的聚类个数 k，一般要用若干个 k 去试验，哪个 k 最后得到的距离平方和最小，就认为哪个 k 是最佳的聚类个数。"徐教授回答说。

李部长问道："徐老师，k-Means 算法第（3）步中的聚类中心是怎么计算的？"

"很简单，将已聚集的点的均值作为新的聚类中心。这正是这种聚类算法成为 k-Means 算法的原因。如果以各聚类均值点最近的点为聚类中心，其他步骤不变，则 k-Means 算法就变为 k-Medoids 算法了。"徐教授回答道。

徐教授突然冒出了个 k-Medoids 算法，又被李部长的双核大脑捕捉到了："徐老师，k-Medoids 算法只是对 k-Means 算法作了个小小的改变，这样有什么作用呢？"

徐教授笑了笑，说："k-Medoids 算法用簇中最靠近中心的一个对象来代表该簇，而 k-Means 算法用质心来代表簇。可见 k-Means 算法对噪声和孤立点数据非常敏感，因为一个离群值会对质心的计算带来很大的影响。而 k-Medoids 算法通过用中心点来代替质心，可以有效地消除这种影响。"

听徐教授这么一解释，李部长又大发感慨："真是小改变，大作用啊！"

马处长觉得他们电力行业对数据挖掘有迫切的应用需求，非常关注算法的应用效果，又问道："k-Means 算法的应用效果怎么样？"

徐教授："**当结果簇是密集的，而簇与簇之间区别明显时，k-Means 算法的效果较好。对于大规模数据集，该算法是相对可扩展的，并且具有较高的效率。**"

李部长不仅脑子转速高，而且善于从反面思考，他又提出了一个问题："徐老师，k-Means 算法和 k-Medoids 算法有哪些不足呢？"

徐教授对答如流："首先，k-Means 算法和 k-Medoids 算法只有在簇数据点的平均值有定义的情况下才能使用。这可能不适用于某些应用，例如涉及有离散属性的数据。"

还没有等徐教授的"其次"出口，一直只听不说的华润超市的万总，被徐教授的这句话触动了，道出了他们数据挖掘时遇到的问题："k-Means 算法和 k-Medoids 算法一般适用于连续变量，而对于离散属性的对象，例如两本书，A=（小说，英文，1/32 开本，浙江大学出版社），B=（计算机图书，中文，1/16 开本，清华大学出版

社），就无均值可言，当然无法使用这两种算法。那么，对于含有离散属性数据的聚类问题怎么办呢？”

徐教授：“为了就解决这类问题，人们对 k-Means 算法进行改进，出现了很多它们的变种，例如，‘k-模’算法用‘模’代替簇的平均值，用新的相异性度量方法来处理分类对象，用基于频率的方法来修改聚类的模。而 k-Means 算法和 k-模算法相结合，用来处理有数值类型和分类类型属性的数据，就产生了‘k-原型’算法。”

听了徐教授的回答，万总非常高兴：“k-模算法和 k-原型算法对我们可太有用了。徐老师，您就详细给我们讲讲 k-模算法和 k-原型算法吧！”

徐教授看了看手表，说道：“按教学计划，这部分是大家课后学习内容，时间不多了，我也就不讲了。你们下去自己看看，有问题咱们一起讨论。”

万总感到有些遗憾，勉强说：“好吧，我们课余时间再与您讨论。徐老师，对不起，刚才您说到 k-Means 算法和 k-Medoids 算法的不足时，我冒昧打断了您的话，您只说了首先，那其次呢？”

徐教授：“其次，这两种算法不适用于发现非球状的簇。原因是这类算法使用距离来描述数据之间的相似性，但是，对于非球状数据集，只用距离来描述是不够的。”

“那遇到非球状的聚类问题可怎么办呢？”万总问道。

徐教授答道：“对于这种情况，要用密度来代替相似性设计聚类算法，这就是基于密度的聚类算法即 Density-based Method。基于密度的算法从数据对象的分布密度出发，把密度足够大的区域连接起来，从而可以发现任意形状的簇，而且此类算法还能够有效去除噪声。常见的基于密度的聚类算法有 DBSCAN、OPTICS、DENCLUE 等。”

李部长已经沉默了好长时间，他担心万总又有什么问题影响徐教授的教学进度，赶紧插话道：“徐老师，您刚才说还有一种层次方法，这种聚类方法的思想……”

徐教授：“好，**我现在就介绍一下层次方法即 Hierarchical Method 的基本思想。这种方法按数据分层建立簇，形成一棵以簇为节点的树。如果自底向上进行层次聚集，**

则称为凝聚的（Aggalomerative）层次聚类；如果自顶向下进行层次分解，则称为分裂法（Divisive）的层次聚类。”

徐教授润了润嗓子，继续讲道：“凝聚的层次聚类首先将每个对象作为一个簇，然后逐渐合并这些簇形成较大的簇，直到所有对象都在同一个簇中，或者满足某个终止条件。分裂的层次聚类与之相反，它首先将所有的对象置于一个簇中，然后逐渐划分为越来越小的簇，直到每个对象自成一簇，或者达到了某个终止条件，例如达到了某个希望的簇数目，或两个最近的簇之间的距离超过了一定的阈值。”

李部长一直认真听着，不断地点头表示他明白了层次聚类的思想。随后，他提问：“徐老师，层次聚类算法有什么缺点？”

徐教授：“层次方法可以在不同粒度水平上对数据进行探测，而且容易实现相似度量或距离度量。但是，单纯的层次聚类算法的终止条件含糊，而且执行合并或分裂簇的操作不可修正，这很可能导致聚类结果质量很低。另外，由于需要检查和估算大量对象或簇才能决定簇的合并或分裂，所以这种方法的可扩展性较差。因此，通常在解决实际聚类问题时要把层次方法与其他方法结合起来。层次方法和其他聚类方法的有效结合可以形成多阶段聚类，能够改善聚类质量。这类方法包括 BIRCH、CURE、ROCK、Chameleon 算法等，它们是如何对层次聚类方法进行改进的、具有什么特点这里不再赘述，大家课后参阅教材。”

其实，李部长对这些经典的聚类算法在他主持硅钢纵条纹质量控制问题的数据挖掘方法研究项目时已经比较熟悉了，他一直在等待徐教授讲解其发明的、享誉国际的聚类算法——视觉聚类算法。

他看了一下手表，过二十多分钟就要下课了。于是，李部长迫不及待地说：“徐老师，您刚才讲了这么多聚类方法，我发现它们有一个共同的缺点，就是算法无法回答数据对象到底可以聚集为多少类，据说你们研究团队发明了一种视觉聚类算法，很好地解决了这一问题。我们几个人昨天晚上还打赌，我说您今天肯定会讲视觉聚类算法，可都快下课了，您根本没有提及‘视觉’两字。我们都等不及了，您还是让我们大家欣赏一下视觉聚类的神奇魅力吧！”

说到**视觉聚类算法**，徐教授脸上露出了会心的微笑。

虽然都连续讲了快两个小时了，他脸上的倦意好像一下子飞到了九霄云外，一个洪亮的声音激荡着教室的每一个角落："同学们，我并不是为我们的视觉聚类算法得到国际上的高度评价而沾沾自喜，作为一个科技工作者，最感到自豪的莫过于他的研究成果在国内外得到广泛应用，为社会的文明、进步作出贡献。"

李部长激动了："徐老师，您就给我们介绍几个视觉聚类算法的典型应用吧！"

"好的，请看大屏幕。视觉聚类算法是基于我们所建立的尺度空间理论建立的，运用这种算法可以对卫星传回的原始图像进行分析，把具有相似属性的事物聚到同一簇中，例如将其用于香港地区地表高精度遥感图像聚类、混杂遥感图像中线状目标如地震带、高速公路、机场跑道等目标识别等。"

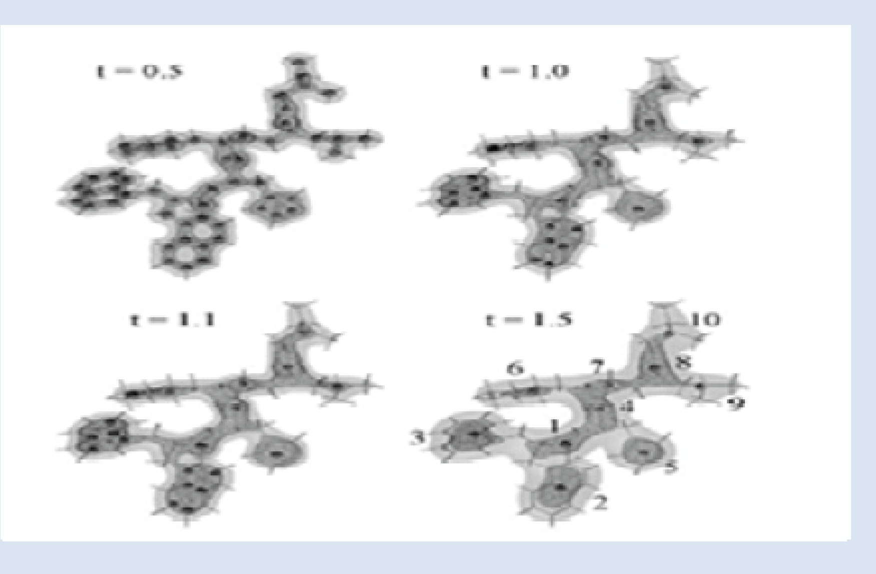
李部长听到这里，激动得跳了起来："徐老师，看来视觉聚类算法有可能用于我们板材表面条纹、夹杂、重皮等质量问题的自动检测，我们试试吧！"

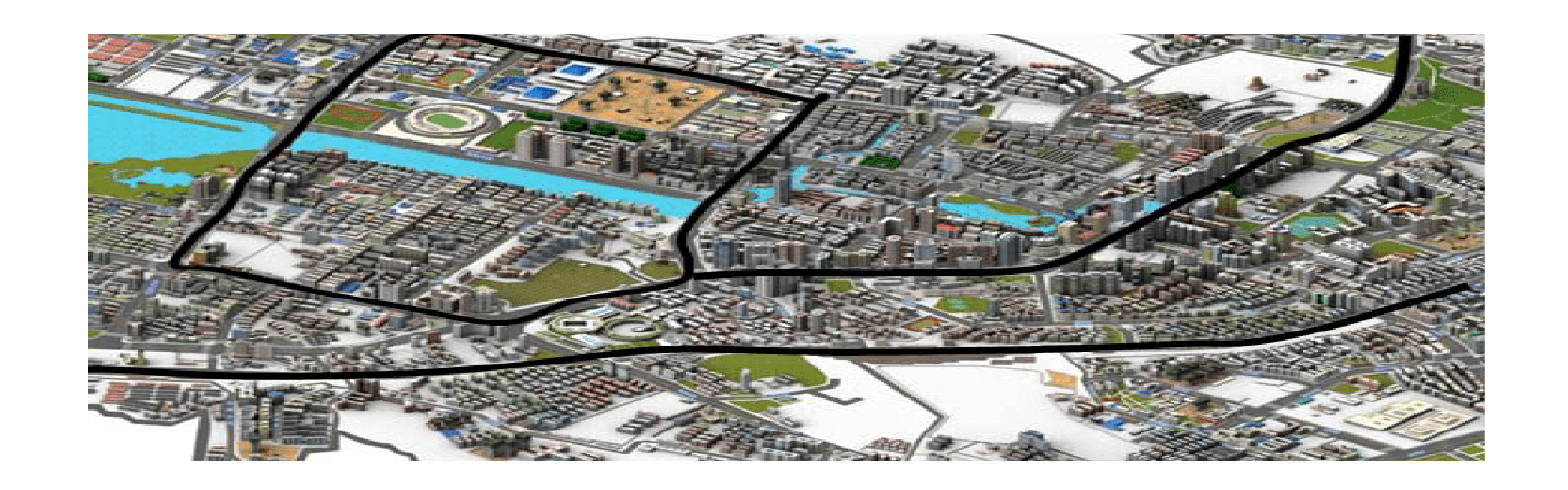
徐教授接着说："李部长的联想很丰富呀，国内外不少公司已经将这种方法试验过了。美国乔治亚大学 Lan 小组、美国马里兰大学 DeMenthon 小组、中科环境与地理信息重点实验室等先后将视觉聚类算法用于地理数据的图像处理，还有比利时 Namur 大学著名的化学家 Leherte 教授所领导的实验室将视觉聚类算法应用到生物计算，进行胃蛋白酶配合体的匹配、分子电流密度函数、蛋白质分子的结构表达等研究。"

蛋白质分析

已得到广泛应用：

地理数据分析(美国乔治亚大学Lan小组);
图像处理(美国马里兰大学Dementhon小组);
蛋白质分析(比利时Namur大学Leherte小组);
中科环境与地理信息重点实验室GAMAX系统;





马处长："徐老师，视觉聚类算法可太有用了，真棒！"

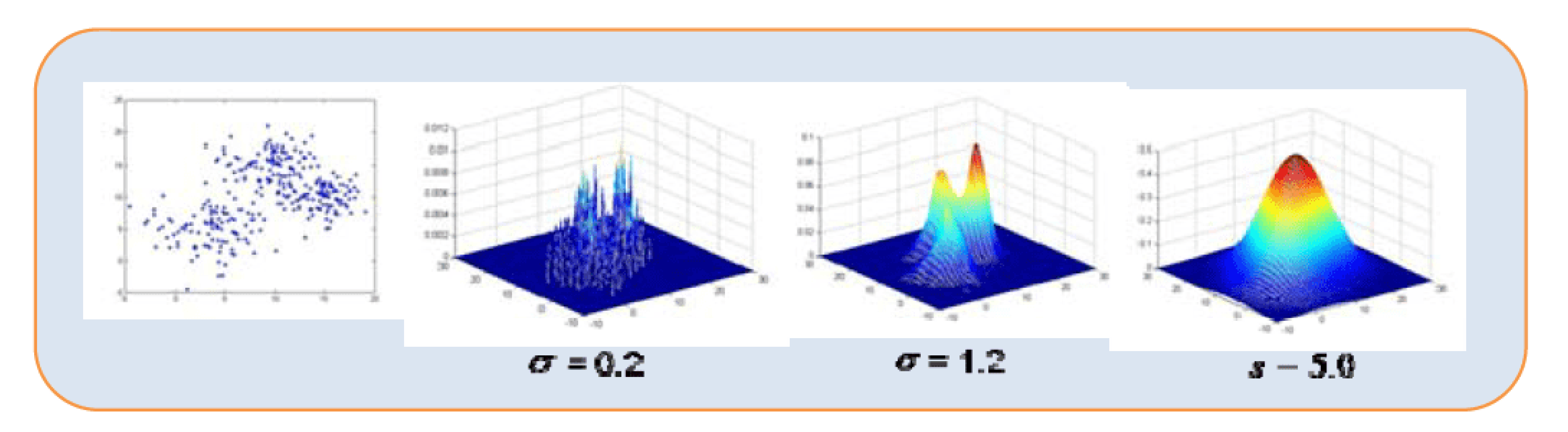
徐教授非常高兴："不谦虚地说，视觉聚类算法确实有其独到之处，其基本思想非常独特：将数据集看作图像，将数据建模问题看作认知问题，通过模拟认知心理学的格式塔原理与生物视觉原理解决问题。"

"且慢且慢，什么是格式塔原理？"李部长打断了徐教授的话语。

徐教授翻动了一下 PPT："很简单，格式塔原理就是物体的整体是由局部特征组织在一起的认知原则，请看屏幕。"



"我们将**相似率**、**连续率**、**闭合率**、**近邻率**和**对称率**作为聚类的基本原则，模拟人的眼睛由近到远观察景物的过程设计算法进行聚类。随着人由近及远，也就是观察尺度由小变大，所看到的景物层次会逐渐变化，实际上这就是一个聚类的过程。"徐教授边说边翻了一下 PPT。



李部长听得如醉如痴，看着 PPT 上视觉聚类的示意图，突然，他冒出了一个新的问题："徐老师，我明白了，在近处，所聚的类会很多，在远处，所聚的类会很少，在很远处，所看到的东西就成为一个类别了。您说，到底聚为多少类最为合适呢？"

徐教授点了点头："李部长的双核脑袋就是转得快，一下子问到了视觉聚类的关键。随着尺度 σ 由小变大，聚类的个数在发生变化，但会出现尺度 σ 在很大范围内变化、而聚类的个数却稳定不变的情况。这个聚类个数存活周期最长，它就是最佳的聚类个数！"

"太妙了，视觉聚类理论通过引进类的生存寿命概念，给出了类的认知定义，解决了聚类有效性问题。数学上严格证明了结构的因果性即类的演化单调性，由此形成了尺度空间聚类的一般性理论框架。"李部长流利地对视觉聚类进行了总结。

徐教授对李部长的话感到纳闷："李部长，你不是做数据挖掘研究的，不可能给出这么深刻的总结吧！"

李部长笑了笑："嘿嘿，这是我从网上看到有人对视觉聚类方法的评价。"

下课铃响了，徐教授边合上电脑边说："聚类方法我们就简单学习到这儿，下一节课咱们一起讨论数据挖掘非常重要的内容——预测。"

1.5.3　预测

这一节要讲预测，学员们兴趣盎然，早早地来到教室。

徐教授走上讲台：“今天我们一起学习数据挖掘的预测方法。”

他的话刚一停顿，就被马处长打断了：“徐老师，税务局的姚局长一直研究周易预测，整天给我们叨叨他料事如神，数据挖掘预测与周易预测有什么不同，哪个更厉害？”

没有想到课堂上会有人提出这样的问题，徐教授灵机一动说：“姚局长，那你就先给大家以最精辟的语言介绍一下周易预测吧！”

姚局长站起来，挠着头：“其实周易预测也是一门科学，马处长、李部长这些人不懂还妄加评论，老是批判我。徐老师，您给了我机会，我得给周易预测正名！”

姚局长越说越激动，徐教授示意他坐下慢慢说。

“周易是建立在阴阳二元论基础上，对天地万物进行性状归类（天干地支五行论），精确到可以对事物的未来发展做出较为准确的预测。周易灵验的预测，千百年来流传，充分证明其具有强大的生命力。其实世俗对周易一直存有误解，比如从迷信的角度去解读它。历史上有许多学者为其正名，他们认为周易理论依据的是万事万物的相似性、关联性和全息性原理。这三个原理已被现代科学所证实，希望人人都能理解，千万不要挖苦讽刺。”姚局长一边说，一边向马处长和李部长投去挑战的目光。

徐教授发现马处长准备站起来反击，急忙以手势示意他坐下。“其实自古以来，确实有太多的伪周易玷污了科学的周易。姚局长和马处长实际上都是科学周易阵线的斗士，但你俩却内讧起来了！”

马处长马上反应了过来：“说来也是，姚局长高举科学周易的大旗，我扛着反击伪科学周易的旗帜，我们本该就是一家人！”说着，马处长将手伸向了姚局长。

这时，姚局长又站了起来，大胆地讲到：“实际上，我们要辩证地看待周易，要以批判继承的观点对待周易。周易在一定程度上揭示和描述了宇宙万事万物运动变化发展的内在规律。如果万事万物不存在相似性、关联性和全息性，周易预测就是不可能的。全息性是周易预测所依据的又一重要原理，科学已经证明了全息性的

存在。”

李部长也搭上了话：“是的，美国科学家做过这样的实验，用一架特制的全息照相机对一棵树苗进行拍照，拍到了一棵大树的照片，后来这棵树苗长大以后正好和这棵大树的照片相吻合。”

听到李部长也开始支持他了，姚局长更起劲了：“考古工作者对一颗牙齿进行化验，得出了古人的身高等许多数据。法医工作者对一根毛发进行化验，得出了死者或者罪犯的许多特征。这说明事物的某一局部包含了其整体的信息。这就是现代科技所证实了的全息论。所以其预测的理论根据是科学的，几千年的实践检验已经证明了这点。”

用余光瞄了瞄同排的专注倾听的学员后，姚局长受到鼓舞，接着说：“可是现在有人硬要把周易预测说成是迷信，那是既不懂周易又不懂科学的表现，是很浅薄的，还有人认为周易很神秘，科学解释不通，这也是不懂科学的表现。周易本身是科学，古老的周易与现代科学是相通的，是血脉相承的。”

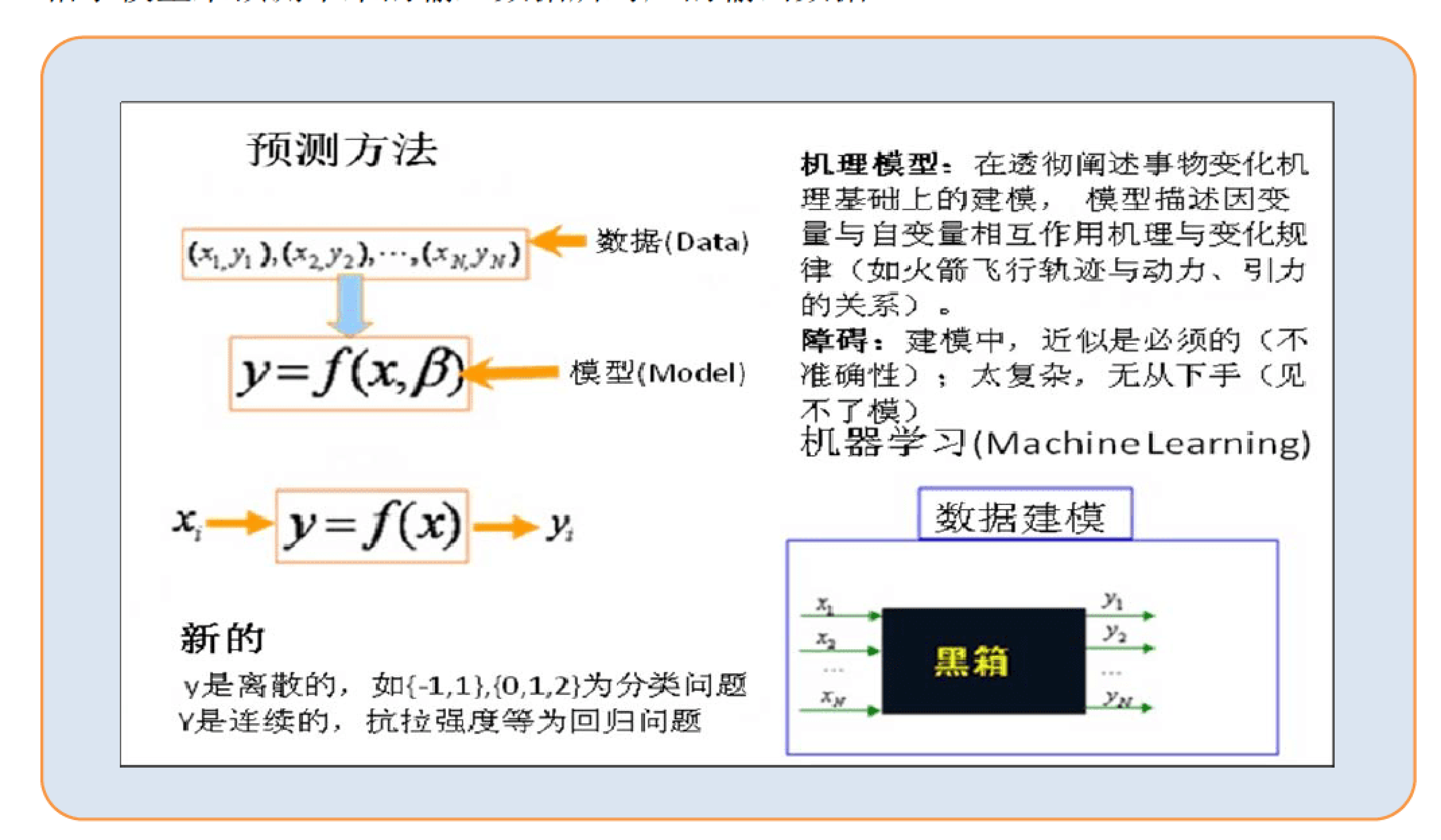
姚部长的一到段话，激起了全班一阵热烈的掌声，结束了 EMBA 班自开学以来对周易的激战。

徐老师觉得是引进数据挖掘的预测概念的时候了，于是说：“数据挖掘的预测是周易预测的继承与发展。周易预测首先要了解事物的属性即状态信息，在利用以往已经获得的事物间的相似性和关联性，对事物的未来状况作出判断。实际上这种相似性和关联性就是对历史事件的学习而积累的经验。而**数据挖掘预测则是通过对反映了事物输入与输出之间的关联性（即内在规律的历史数据）的学习，得到预测模型，再利用该模型对未来数据进行预测的过程**。”

马处长对徐教授所讲的内容感到疑惑不解，大声请求道：“徐老师，您讲得太深奥了，能不能再具体地描述描述数据挖掘预测的过程？”

徐教授将 PPT 翻到新的一页，说道：“数据挖掘预测的基本原理是黑箱子模型，即将事物输入与输出之间的关系不管其多么复杂，均当做一个黑箱子，以往的输入、

输出数据是这个黑箱子内复杂规律的反映。通过数据挖掘的机器学习方法，建立黑箱子模型来预测未来的输入数据所对应的输出数据。”



“慢点慢点，徐老师，什么是机器学习？”马处长捕捉到了一个新名词，急忙问道。

徐教授早已预料到有人会问这样的问题，不紧不慢地说道：“假定事物的输入、输出之间存在一种函数关系 $y=f(x,\beta)$，其中 x 是待定参数，$y=f(x,\beta)$ 称为学习机器。通过数据建模，由历史输入输出数据学习得到参数 β，就确定了的具体表达形式 $y=f(x,\beta)$，于是便可以对新的 x 预测 x 了。这样的过程称为机器学习。”

“徐老师，我只听说过数学建模，您刚才提到数据建模是什么意思？”姚部长也提出了一个问题。

“数据建模就是基于数据建立数学模型，它是相对于基于物理、化学和其他专业基本原理建立数学模型（即机理建模）而言的。对于预测来说，如果所研究的对象有明晰的机理，可以依其进行数学建模，这当然是最好的选择。但是，我们经常会遇到很多实际问题，如社会学问题、金融问题、复杂工业过程问题和生物医学问

题等，不适合以某种机理来描述，从而无法进行机理建模。但如果积累有足够的历史数据，这时，数据建模就可大显身手了。”

受徐教授的启发，学员们纷纷谈论其本行业的情况。

李部长深有感触地说：“冶金工业是极其复杂的流程化生产过程，各个工序对产品质量都有影响，尤其是产品表观质量问题（如冷轧板重皮、夹杂、侧翻和硅钢纵条纹等缺陷）根本无法建立机理模型。不过，冶金生产自动化程度很高，数据积累非常丰富，为数据建模提供了良好的基础。”

李部长的话也引起了马处长的共鸣：“在我们电力行业，设备状态及寿命评估、负荷预测、电力暂态稳定性分析、电力系统规划等诸多问题都难于进行机理建模，机器学习可以发挥重大作用了。”

铁路局的高副局长也开了口：“在我们铁路部门，高铁的轨道检测、交通流量预测、铁路票价制定、调度优化等，均可以用机器学习的方法解决啦！”

税务局赵局长也忍不住了：“好啊，税务稽查也有数据挖掘这把利器了！”

航天研究院的黄主任接着说：“说起机器学习，我这里有个非常典型的实例跟大家分享，就是关于劳动定额的预测。以某飞机零部件生产加工为例，通过分析历史数据中的加工宽度、加工直径、加工深度和劳动定额之间的关系，最终建立起各加工尺寸和劳动定额的 BP 神经网络回归预测模型。经过对模型的效果分析评估，我们将此模型固化应用在实际生产中几个月后，发现此模型预测准确率高达 99.21%，帮助企业节省了大量的收集数据的经济和时间成本。更具现实意义的是，将得到的劳动定额制度在企业的生产中组织贯彻，并采取有关的技术组织措施，如竞赛、技术培训、动作分析、定额考核等，能帮助职工达到和不断突破现行劳动定额。根据职工完成定额的情况进行分析，管理者亦能发现定额管理中存在的问题并加以解决。”

工行的张行长显得非常平静，慢条斯理地说：“其实，我们已经开始尝试利用机器学习的方法进行信用评价、贷款风险评估和反洗钱等工作，希望徐教授和其他

学员不吝指教。”

移动公司梁总显得有点得意，喜形于色地说：“我们公司两年前就开始应用数据挖掘解决电信业面临最紧迫的四大问题：市场分群、精确营销、新业务响应和客户流失分析等。这四大问题最本质的还是预测问题，我们已经总结出了比较成功的解决方案，有机会邀请徐教授给我们指导指导。”

“……”

“好的，大家都讲了很多了。预测未来趋势和行为，使得行动目标更具有前瞻性，并作出具有知识驱动的决策，是每一个行业的共同希望，但愿数据挖掘的机器学习方法能使大家以后的工作如虎添翼。”徐教授总结道。

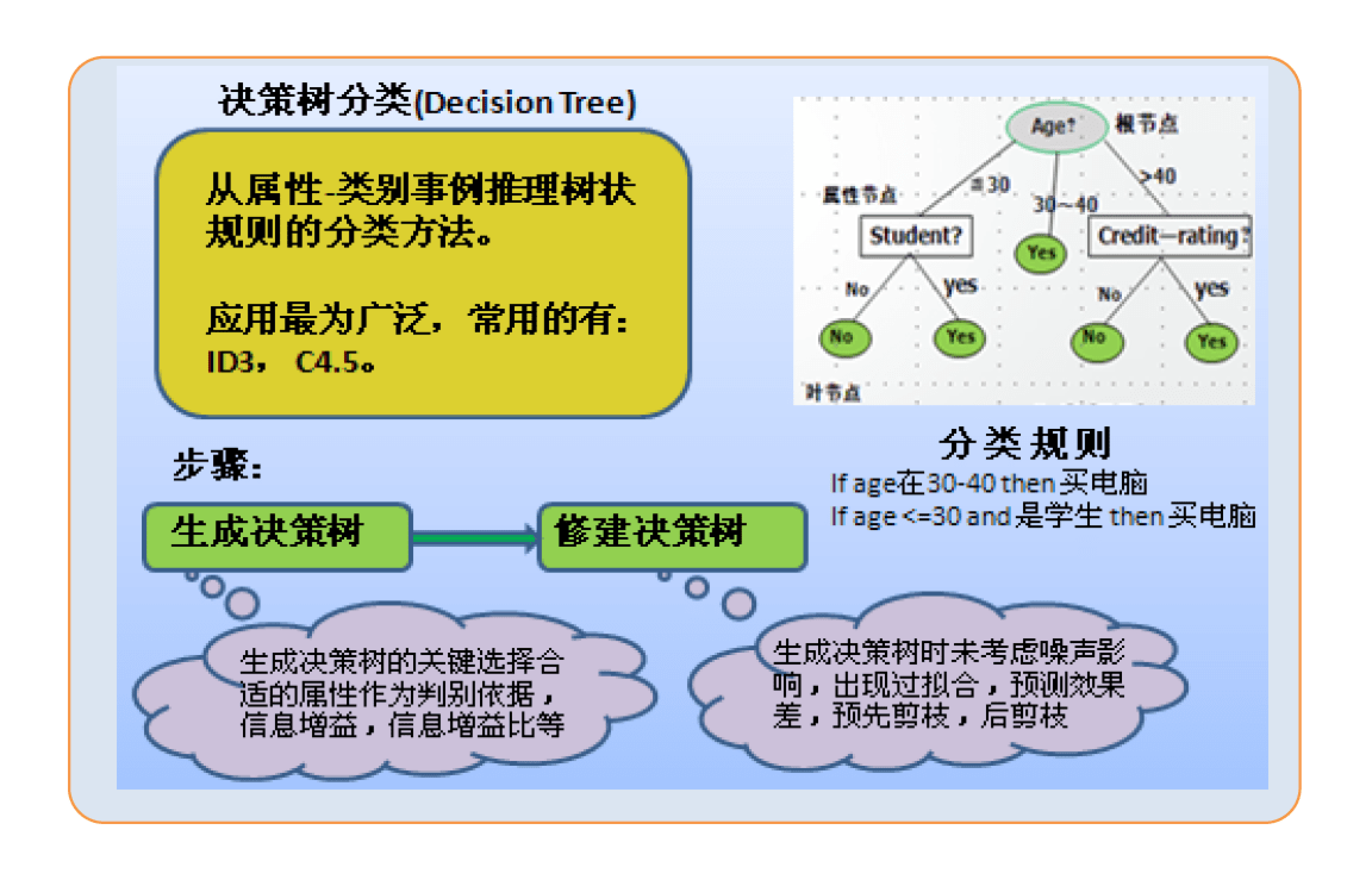
马处长估计徐教授下面要讲具体的机器学习方法了，急忙建议说：“徐老师，机器学习的数学模型和求解算法很多，而且新方法层出不穷，在应用中让人难以选择，您就给我们讲解一些实用而先进的方法吧。”

徐教授没有回答马处长的话，只是轻轻点了一下鼠标，几种典型的机器学习方法在屏幕上跃然而入。

- 决策树方法
- 人工神经网络
- 支撑向量机
- 正则化方法

（1）决策树方法

徐教授的 PPT 又翻开了新的一页，他将光笔指向屏幕上的树状图，讲道：“所谓决策树就是一个类似流程图的树型结构，树的最高层结点就是根结点，树的每个内部结点代表对一个属性（取值）的测试，其分支代表测试的每个结果，而树的每个叶结点代表一个类别。从根节点到叶子节点的每一条路径构成一条‘IF…THEN…’分类规则。”



李部长凝视着大屏幕上的决策树，明白了其中的奥妙，不禁说道：“决策树方法实际上就是通过一定的评判策略判定哪一个属性对分类最为重要，就将其作为根节点，然后再判断余下的节点中最重要的节点，直到叶子节点。”

“好，理解得还比较透彻。不过，李部长，什么样的节点才可以标注为叶子节点呢？”徐教授问。

李部长吱吱唔唔：“好像有三种情况……”

“对，符合以下三个条件之一的节点就可为叶子节点：（1）节点的样本集合中所有样本都属于同一类；（2）节点的样本集合中所有的属性都已经处理完毕，没有剩余属性可以用来进一步划分样本，这时候采用子集中多数样本所属于的类来标记该节点；（3）节点的样本集合中所有样本的剩余属性取值完全相同，但所属类别却不同，此时用样本中多数类来标示该节点。”

徐教授接着说：“决策树算法的典型代表是 ID3（Interactive Dicremiser version 3）算法，它是由 Quinlan 等人于 1986 年提出的，是当时机器学习领域中最有影响力的算法之一。其核心思想是在决策树的构建过程中采取基于信息增益的特征选择策略，即选取具有最高信息增益的属性作为当前节点的分裂属性，使得对结果划分中的样本分类所需要的信息量最小。以此构造与训练数据一致的一棵决策树，从而保证了决策树具有最小的分支数量和最小的冗余度。”

李部长：“ID3 算法思想简单，并且由其构造的决策树对样本的识别率比较高。在实际应用中，ID3 算法有什么不足之处吗？”

徐教授按了一下光笔，并说：“请看大屏幕 ID3 算法的缺点主要表现在以下几个方面。”

ID3 算法的不足之处

（1）ID3 算法在搜索过程中不能回溯重新考虑选择过的属性，从而收敛到局部最优解而不是全局最优解；

（2）信息增益的度量偏袒于属性取值数目较多的属性，这不太合理；

（3）ID3 算法只能处理离散值的属性，不能处理连续属性；

（4）当训练样本过小或者包含有噪声的时候，容易产生过度拟和（Overfitting）现象。

马处长看着屏幕，问道：“徐老师，那怎样改进 ID3 算法呢？”

徐教授回答道：“针对 ID3 算法的不足，Quinlan 于 1993 年提出了 ID3 的改进方法——C4.5。与 ID3 相比，C4.5 主要在以下几个方面作了修改，并且引进了新的功能：用信息增益比率作为选择标准，弥补了 ID3 算法偏向于取值较多的属性的不足；合并连续属性的值；可以处理具有缺少属性值的训练样本；运用不同的剪枝技术来避免决策树的过拟合现象；K 次交叉验证等。”

李部长又问：“徐老师，我们在使用决策树算法进行分类时，有时会出现过拟合现象，这是怎么回事呢？”

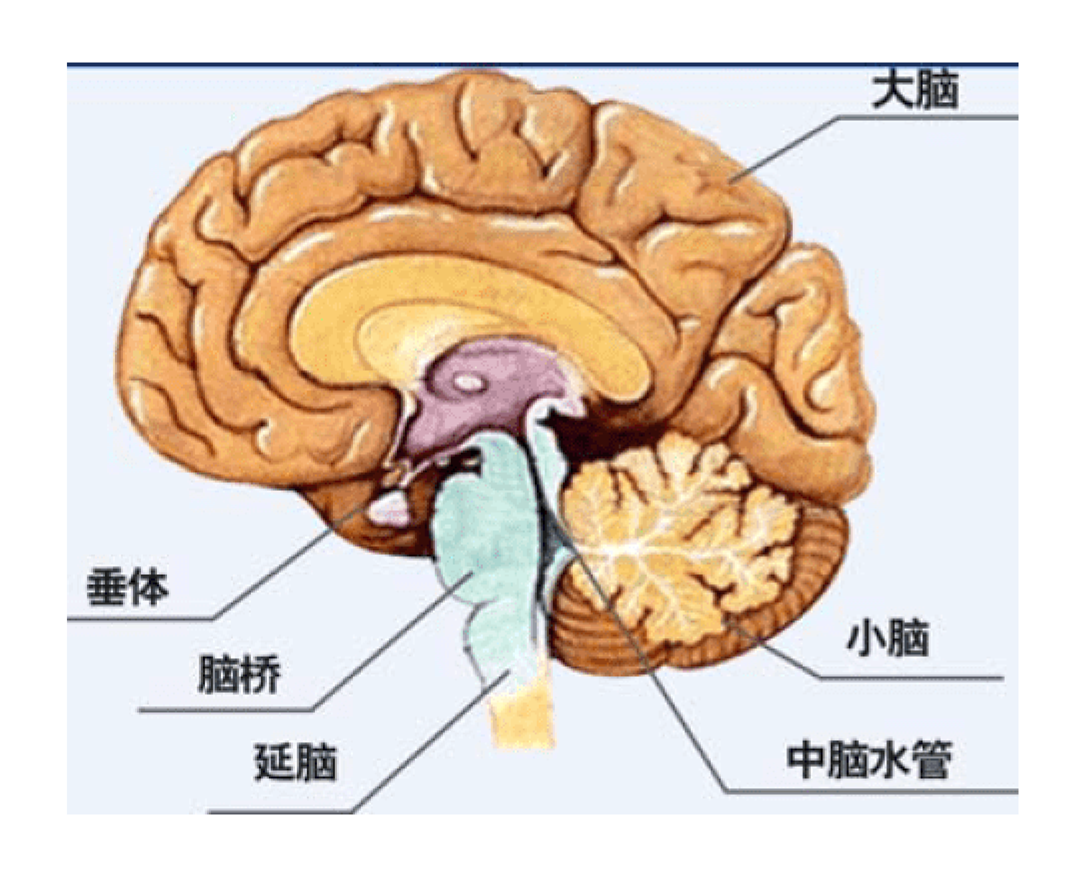
徐教授不厌其烦："基本的决策树构造算法没有考虑噪声，因此生成的决策树可以完全与训练数据拟合，也就是说，对训练数据的测试准确度可以达到 100%。但是在有噪声的情况下，完全拟合将导致'过拟合'的结果，即对训练数据的完全拟合反而导致对新数据的预测能力下降。这是因为当训练数据集合包含噪声时，决策树在生成的过程中为了与训练数据一致，必然生成了一些反应噪声的分支，这些分支不仅可能在新的决策问题中导致错误的预测，而且增加了模型的复杂度。"

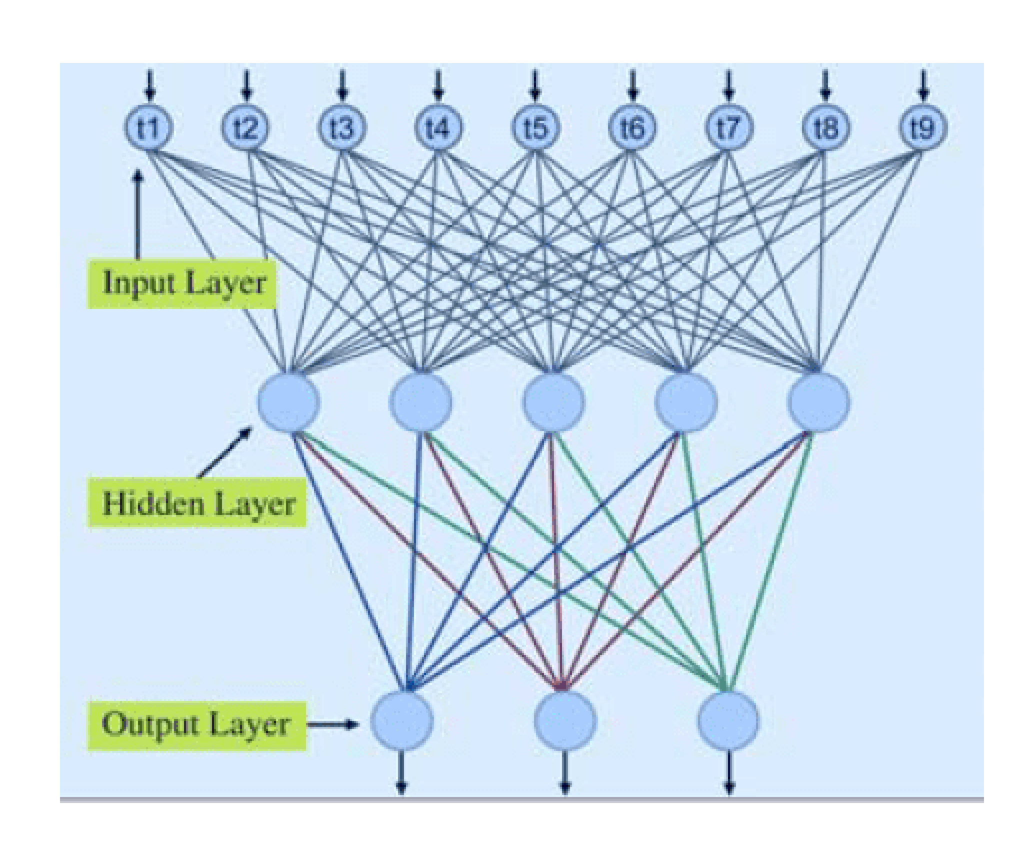
马处长也问道："那怎么避免过拟合现象呢？"

徐教授："解决决策树生成过程中的过拟合问题的方法主要是对决策树进行剪枝。剪枝是一种克服噪声的技术，它有助于提高决策树对新数据的准确分类能力，同时能使决策树得到简化，使其更容易理解，加快分类速度。剪枝策略可分为预剪枝（pre-pruning）和后剪枝（post-pruning）两种。预剪枝主要是通过建立某些规则限制决策树的充分生长，后剪枝则是等决策树充分生长完毕后再剪去那些不具有一般代表性的叶节点或者分枝。尽管前一种方法可能看起来更直接，但是后一种方法在实践中更成功。因此在实际运用中更多的采用后剪枝技术。"

（2）人工神经网络

徐教授的 PPT 翻到了新的一页，一个人脑结构图和密密麻麻的结构图跃入屏幕。他用光笔指着图讲到："人工神经网络，Artificial Neural Networks，简写为 ANNs，是对人脑若干基本特性的抽象。它由大量神经元通过丰富的连接构成多层网络，用以模拟人脑功能。"





“还能模拟人头脑的功能，这么厉害？”有人感到不可思议。

“实际上，神经网络只是个不依赖于模型的自适应函数估计器，可以实现任意的函数关系。”徐教授补充道。

“那也就是说，人工神经网络是一种机器学习方法，也可以对求解分类和回归问题进行预测。”马处长道出了自己的理解。

“更有用的是，定量或定性的信息都可贮存于网络内的各神经元中。也就是说，它可以同时处理定量、定性知识。而且网络有很强的稳定性和容错性。”徐教授补充道。

（3）支撑向量机

“支撑向量机，Support Vector Machines，简称 SVM，是 20 世纪 90 年代 Vapnik 等人根据统计学习理论中结构风险最小化原则提出的一种机器学习方法。”徐教授说。

“它用来解决分类问题还是回归问题？”马处长问道。

“既可以求解分类问题，也可以用于回归问题。但是，起初是从分类问题建模的，后来又拓展到求解回归问题。”徐教授回答道。

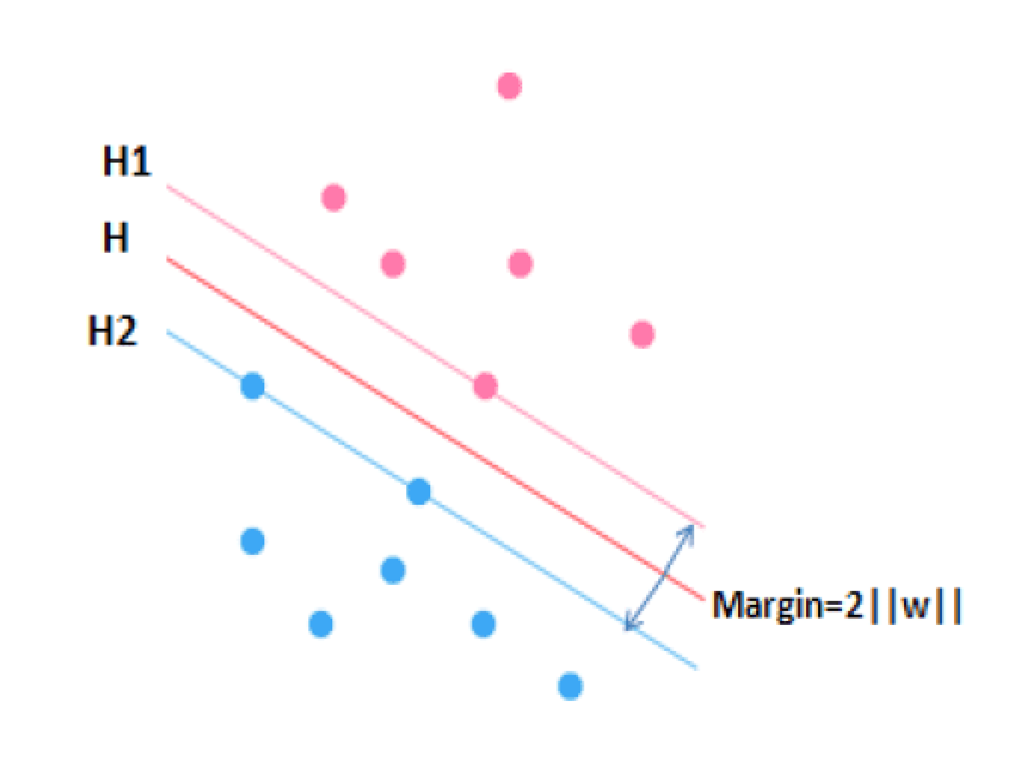
“徐老师，听我们单位去年来的博士小施说，支撑向量机用处太大了。您就深入浅出地介绍一下其建模原理吧。”

“支撑向量机是从线性可分的二分类问题开始建模的，再逐步向线性不可分问题、非线性问题深入，最后推广到线性和非线性回归问题建模。”

“那您就从最简单的、线性可分的二分类问题讲起吧。”马处长建议说。

“好吧。请看屏幕，对了，不是手机屏幕，是投影屏幕。”徐教授这么一说，玩弄手机的学员不好意思地将手机藏了起来。

“图中，方形点和圆形点代表两类样本，H 为分类线，H1、H2 分别为通过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫做分类间隔（margin）。所谓最优分类线就是要求分类线不但能将两类正确分开，而且使分类间隔最大。推广到高维空间，最优分类线就是最优分类面。”



“之所以要求得分类间隔最大的最优分类面是为了对未来的新样本预测得更准确。”李部长早已对 SVM 很熟悉了，补充说。

“对这一问题，前苏联人 Vapnik 等人于 1995 年建立了以分类间隔最大化为目标，以分类面将样本全部区分正确为约束条件的二次优化模型。”徐教授说。

“对这个模型进行怎样的改变，就可以处理线性不可分问题？”马处长动了脑筋。

“只要将约束条件放宽为‘允许分错’就行了。”徐教授回答说。

“对于分类面为曲面的分类问题，怎么处理？”马处长又问。

“通过引进该函数，进行非线性变换，将输入数据变换到一个高维空间，在这个高维空间里，原来低维空间的曲面，变成了平面，就可求解最优分类超平面的方法了。”徐教授回答道。

“妙，实在是妙！复杂的非线性分类问题线性化了。”马处长感慨道。

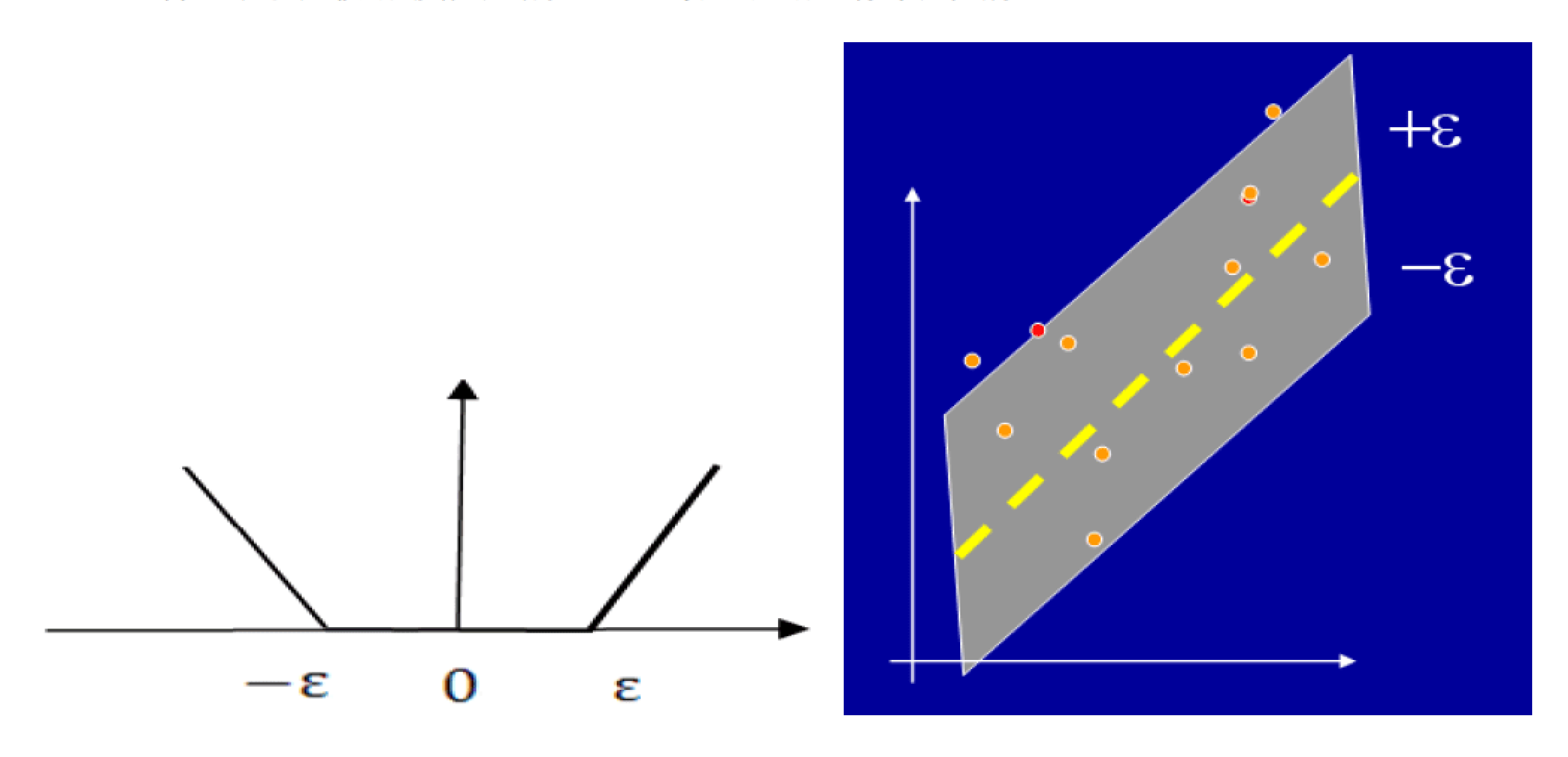
“不过，上研究生的时候，跟着导师搞课题，曾经遇到过二次规划问题，几十个变量、3000 多条数据，双核电脑就跑不动了。支撑向量机模型求解可能会遇到麻烦。”马处长问题连连。

徐教授回答道："曾经，这确实是个问题，1998 年，微软研究院的 John C. Platt 提出了最快的求解二次规划的 SMO 算法，这一问题迎刃而解。"

马处长又问道："后来怎么将这向量机扩展为可求解回归问题？"

"这就更加奥妙了，'不敏感损失函数'功不可没。"徐教授叹道。

"什么是不敏感损失函数？"马处长感到莫名其妙。



徐教授回答道："损失函数就是衡量回归结果与真实值相差大小的一种函数。不敏感损失函数定义为绝对误差，即回归结果与真实值之差的绝对值，小于一定的值 ε 时，就认为回归函数对预测没有造成损失，否则损失就为其绝对误差。"

徐教授接着道："通过不敏感损失函数，将样本点分成了绝对误差小于 ε 和大于 ε 两类，这样就可以用分类方法建立支撑向量机回归模型了。"

"妙，妙，更加的妙！"马处长道。

"那么支撑向量机有什么优势呢？"台下有人问。

"它在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势。SVM 建立在计算学习理论的结构风险最小化原则之上，具有简洁的数学形式，能进行直观的几何解释，并具有良好的泛化能力，避免了局部最优解，且需要人为设定

的参数少，便于使用，为小样本机器学习提供了一种新方法。”徐教授说。

（4）正则化方法

“现在，我们简单了解一下机器学习的一种新方法——正则化方法。”徐教授说道。

“神经网络和支撑向量机方法不是很好嘛？”马处长提出了疑意。

“是的，神经网络和支撑向量机方法是应用比较普遍的机器学习方法。但各种方法都有其优缺点。”徐教授解释道。

“那么神经网络有什么不足？”一学员问。

“神经网络容易陷入局部极小点；易出现‘过拟合’而使得泛化能力较差；而且网络拓扑结构的确定没有成熟的理论指导；神经网络训练代价很高；其解不具有稀疏性和且难以解释。”徐教授如数家珍地说。

“支撑向量机也有很多缺点吗？”那位学员接着问道。

“支撑向量机方法是在机器学习理论指导下专门针对有限样本设计的学习方法，不仅对于小样本问题可以得到最优解，而且 SVM 模型具有很强的泛化能力。更为突出的是 SVM 最终转化为求解一个凸二次规划问题，在理论上可以得到全局最优解，克服了一些传统方法（如神经网络方法）可能会陷入局部极值的不足。虽然支撑向量机与神经网络相比有着明显的优势，当在实际应用中还存在着一些问题，比如对于太大规模的数据集，由于 SVM 要解凸二次规划而使算法效率很低，甚至算法无法进行；SVM 对奇异值的稳健性不高；SVM 的解不具有稀疏性，存在着大量冗余支撑向量等；更令人感到美中不足的是参数没有好的选择策略。这些不利因素限制了 SVM 在一些领域的应用。”

徐教授一口气把支撑向量机的优缺点对比得清清楚楚。

“徐老师，看来您还会给我们介绍更好的机器学习方法。”马处长猜测道。

“近年来，正则化方法得到了机器学习研究者的广泛关注，人们提出了不少满足不同性能要求的基于正则化的框架模型，其典型代表有 Lasso 模型和推广的 Lasso 模型、L1/2 正则化模型及其迭代阈值算法等。”徐教授介绍说。

“这些模型与神经网络和支撑向量机比，有什么优势？”马处长问。

“由于 Lasso 方法用模型系数的绝对值函数作为惩罚来压缩模型系数，使绝对值较小的系数自动压缩为 0，这样得到的模型具有稀疏性，从而同时实现显著性变量的选择和对应参数的估计。”徐教授讲道。

“徐老师，我听您说过，Lasso 模型采用平方损失函数而使其稳健性较差，而且在很多应用场合（例如分类问题）损失函数不宜采用平方损失，这就使得 Lasso 模型的应用受到限制。”李部长回忆道。

“因此我们对 Lasso 模型进行推广，使推广后的 Lasso 模型可以使用其他损失函数，并可应用于回归问题和分类问题。”徐教授回应道。

“Lasso 模型和推广后的 Lasso 模型都属于 L1 正则化模型，用什么算法求解比较好？”李部长又问。

“这两类模型是凸优化问题，有很多算法可以求解，但梯度 Boosting 算法更为实用。”徐教授给出了建议。

“徐老师，L1 正则化模型就具有稀疏性，求解也比较容易，为什么还要建立 L1/2 正则化模型？”李部长接着问道。

“理论研究和实验证实，L1/2 正则化模型解比 L1 正则化模型的解更稀疏，虽然他为非凸优化问题，难以求解，但我们提出了 L1/2 迭代阈值算法，可巧妙而高效地对其求解。”

1.5.4　序列和时间序列

“前面我们讲过分类和聚类是双胞胎兄弟，这节课我们所讲述的序列和关联分析也是双胞胎兄弟。”徐教授说。

“序列是怎么样一个东西呢？”有人问。



“序列就是被排成一列的对象（或事件），这样，每个元素不是在其他元素之前，就是在其他元素之后，元素之间的顺序非常重要。就如电话号码一样，同样的数字但是不同次序代表了很重要的信息。”



“次序有这么重要？”有人嘀咕。

“例如 119 火警，你按成 911，不好意思你打到美国报警电话了，哈哈。”徐教授说。



“嗯，明白了，也就是序列具有了次序属性，对吧？”

“是的，序列与关联关系很密切，所不同的是在序列发现中事件的相关是以次序来区隔，有时候是以时间来区隔。”

“原来这样！那徐老师您给举个具体点的例子吧？”有人提议。



“例如：如果 A 股票在某一天上涨 12%，而且当天股市加权指数下降，则 B 股票在两天之内上涨的机率是 68%。”徐教授说。

“这两个双胞胎的具体区别怎么来总结呢？”

“序列也是发现组合规律的，不过关联中所提到的规律不涉及先后次序，而序列则是有先后次序的。那谁知道时间序列？”徐教授。

“莫非序列元素有了时间属性？”

“是的，对于具有时间属性的序列进行分析，就用到了时间序列分析。时间序列分析是指通过对大量时间序列数据的分析找到特定的规则和感兴趣的特性，从而实现对未来状态的预测。”徐教授说。

“徐教授，那么回归分析和时间序列分析有什么区别呢？”有人问。

“时间序列预测和回归的功能类似，只是时间序列是用历史数值来预测未来数值，是一种特殊的自回归，更多的表现为描述对于过去时刻的观测和相应时刻的随机扰动的记忆性规律。”徐教授解释说。

“徐老师，刚才您说股票数据是一种时间序列，在中国这个政策性市场条件下，时间序列可能发挥不到什么大的作用。但说到它在冶金企业中的巨大作用，我可是深有体会，对于我们这些冶金企业来说，用于抽取烧结过程产生废气的风机是一个关键设备，过去我们常常需要定期停机检修。自从公司对它建立了时间序列模型，很好地预测了它将来的状态，不仅减少了停机成本，而且降低了维修上的费用。”李部长感慨道。

“是的，对于时间序列模型国人还是比较熟悉的，从气象预测到设备的状态检修等都有着成功的应用案例。”徐教授很肯定地说。

1.6　数据挖掘工具

“黑格尔说：存在即合理。”，徐教授用一句名言开始了本节课的内容。

台下一个学员悄悄地说：“没想到徐教授还研究哲学……”

“我最近开始研究盲信号处理，尽管你的声音很小，我还是听见了，谢谢夸奖。”

大家都被逗乐了。

徐教授："存在即合理，由此可以引申出另外一句名言：哪里有需求，哪里就有产品。"

"就是，需求是产品的源动力啊"，大家对这个观点的看法是完全认可。

接着，徐教授便进入今天的正题："由于数据挖掘的强大功能，能为社会创造巨大的经济效益，一些著名大学和国际知名公司纷纷开发相关的软件产品。下面我们来个小摸底，听说过或者使用过数据挖掘软件工具的同学请举个手。"

只见下面的学员中，稀稀拉拉只有 4～5 个人举手。

EMBA 的课桌前面插着每一个学生名字的牌子，字很大、很清楚。老师提问时可直呼其名，同学回答问题时也都使用麦克风，以便教室里的所有人都能够听清楚。对于在课堂上这样暴露身份，并在所有同学注视下大声讲话，老总们开始不太适应，被老师点到名字时总有些紧张。

徐教授指着刚才举手的张经理问道："张经理，给大家说说你了解的数据挖掘工具。"

张经理腼腆地笑道："其实我自己没用过，不过见我们公司技术部小赵使用过数据挖掘软件，听他说那软件是 IBM 的 Intelligent Miner。"

徐教授回应道："对，IBM 的 Intelligent Miner 是 IBM 公司 1996 年推出的数据挖掘产品，包含多种统计方法和挖掘算法，可以进行线性回归、因子分析、主变量分析、分类、分群、关联、相似序列、序列模式、预测、发现关联、发现序列规律、概念性分类和可视化呈现，还可以自动实现数据选择、数据转换、数据挖掘和结果呈现等一系列数据挖掘操作。"

"上我的课，不要紧张，根本没必要'十五个吊桶打水，七上八下'的。刘总，我看刚刚你也举手了，跟大家分享一下你所了解的数据挖掘工具。"

刘总站起来回答道："我接触数据挖掘工具时间比较短，我们部门使用的是 Unica

Model 1。”

徐教授：“刘总，问你一个可能涉及到隐私的问题，你负责你们公司产品的营销活动策划吧？”

刘总说：“是的，徐老师，你比外边那些算命的能掐会算多了。”

姚局长说：“徐教授，莫非您也精读了周易？”

大家都笑翻了，开始更加好奇徐老师是怎么知道的呢。

徐教授：“因为 Unica Model 1 这个软件是一款典型的、针对市场营销和策划行业而研发的软件。”

“原来是这样”，学员们恍然大悟。

徐教授接着说：“Unica Model 1 这个软件很经典，非常畅销。它涵盖了响应模型、交叉销售模型、客户价值评估模型、市场细分模型等，这四部分简直就是这个软件的四大金刚。还有那个同学愿意自告奋勇地给大家讲讲其他数据挖掘工具？”

工行的张行长说：“我对 SAS 软件了解一些，该系统全称为 Statistics Analysis System，最早由北卡罗来纳大学的两位生物统计学研究生编制，并于 1976 年成立了 SAS 软件研究所，正式推出了 SAS 软件。经过多年的发展，SAS 已被全世界 120 多个国家和地区的近 3 万家机构所采用，直接用户则超过 300 万人，遍及金融、医药卫生、生产、运输、通讯、政府和教育科研等领域。”

上海一家钢铁公司的贾总站起来了，补充说道：“我们公司使用的就是 SAS 软件。由于 SAS 系统是从大型机系统发展而来，在设计上也完全针对专业用户，因此其操作至今仍以编程为主，人机对话界面不太友好，并且在编程操作时需要用户最好对其使用的统计方法有较清楚地了解，非统计专业人员掌握起来较为困难。而且 SAS 极为高昂的价格和只租不卖的销售策略使得实力不足的个人和机构只能望而却步。不过，由于其功能强大，我公司专业人员较多，这几年我们不惜巨资每年都在租用该软件。”

徐教授感到很惊讶："咱们这个班果真卧虎藏龙。张行长和贾总回答地非常专业。不知道的人还以为你俩是 SAS 公司的'山寨'销售专家呢。"

贾总笑了笑，不好意思地说："我大学同宿舍的一位同学在 SAS 北京办事处工作，经常来上海推销他们的产品，每次顺便来我这儿蹭酒喝，免不了给我叨叨他们的 SAS，时间长了我就耳熟能详了。"

徐教授也乐了："原来如此！"

徐教授的话音刚落，市统计局程副局长立即站了起来："SAS 太专业了，我们统计分析用 SPSS。"

徐教授："好，那我就简要的向大家介绍一下 SPSS 统计软件吧。1968 年，斯坦福大学三位学生创建了 SPSS 公司，最初定位为'社会科学统计软件包'即 Solutions Statistical Package for the Social Sciences，但是随着 SPSS 产品服务领域的扩大和服务深度的增加，SPSS 公司已于 2000 年正式将其更改为'统计产品与服务解决方案'即 Statistical Product and Service Solutions。其最突出的特点就是操作界面极为友好，输出结果美观漂亮。它将几乎所有的功能都以统一、规范的界面展现出来，使用 Windows 的窗口方式展示各种管理和分析数据方法的功能，对话框展示出各种功能选择项。"

"这么说 SPSS 一定很好用了？"刚才提问的那位学员继续问道。

统计局程副局长深有感触地说："用户只要掌握一定的 Windows 操作技能，粗通统计分析原理，就可以使用该软件进行统计分析或数据挖掘。现在全球约有 25 万家以上用户，分布于通讯、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业。目前 SPSS 是世界上应用最广泛的专业统计软件。"

"徐老师，SPSS 有哪些主要功能？"一个学员问。

徐教授："SPSS 的基本功能包括数据管理、统计分析、图表分析、输出管理等。SPSS 统计分析过程包括描述性统计、均值比较、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、数据简化、生存分析、时间序列分析、多重响应等几

大类。具体每类中又分好几个统计过程，比如回归分析中又分线性回归分析、曲线估计、Logistic 回归、Probit 回归、加权估计、最小二乘法、非线性回归等多个统计过程，而且每个过程中又允许用户选择不同的方法及参数。SPSS 也有专门的绘图系统，可以根据数据绘制各种图形。”

上海钢铁公司的贾总一直认真地听着，终于沉默不住了：“其实 SPSS 公司的真正的数据挖掘产品是 Clementine。它的图形化工作流操作方式使得分析人员能够看到数据挖掘过程的每一步。通过与数据流的交互，分析人员和业务人员可以合作，将业务知识融入到数据挖掘过程中。这样数据挖掘人员就可以把注意力集中于知识发现，而不是陷入技术任务，例如写代码，所以他们可以尝试更多的分析思路，更深入地探索数据，揭示更多的隐含关系。我们公司也有不少技术人员对 Clementine 爱不释手。不过，网上说 2009 年 7 月，IBM 以 12 亿美元现金收购了 SPSS 公司，Clementine 也更名为 IBM SPSS Modeler 了。”

航天研究院的黄主任：“变成 IBM 的软件，那不就更贵了。不过近几年有一款免费的数据挖掘软件 WEKA，异军突起。”

贾总的钢铁公司为世界 500 强企业，财大气粗地说：“管它免费不免费，软件到底好用不好用？”

黄主任从座位上站了起来，细声细语：“WEKA 的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），是一款免费的、基于 JAVA 环境下开源的数据挖掘软件，1993 年由新西兰的 the University of Waikato 进行开发。WEKA 集成了非常多的数据挖掘和机器学习算法，包括分类。回归、聚类、关联规则等方面。2005 年 8 月，在第 11 届 ACM SIGKDD 国际会议上，the University of Waikato 的 WEKA 小组荣获了数据挖掘和知识探索领域的最高服务奖。从此 WEKA 系统得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一。”

徐教授示意黄主任坐下，总结道：“但是，WEKA 算法多的优点对于数据挖掘非专业用户来说反而变成了缺点，用户往往无法判断选择哪些算法适合解决自己的问

题。说实在的，其中不少算法只是科研成果，并不实用。”

最后，徐教授在黑板上写下了一个网址（http://www. datamininglab. Com），并说：“除了上面提到的这些数据挖掘软件外，大家感兴趣的话可以自己光顾这个网站，该网站还提供了许多数据挖掘工具软件的性能测试报告。”

读书笔记

第 2 章　数据挖掘流程

上一节课结束时，徐教授建议让国内不锈钢巨头的品质部李部长与大家分享他们公司数据挖掘的成功经验，向其他学员介绍数据挖掘的流程，李部长欣然答应。今天李部长比平常来的早，而且西装革履，皮鞋锃亮，头发油光可鉴。上课铃声一响，他便健步走上讲台，绝对是大学者风范。

“各位领导，我不是一位数据挖掘的专家，但是，我敢大言不惭地说，我是工业界敢吃‘螃蟹’者之一。今天我只想把我们公司应用数据挖掘技术解决硅钢质量控制难题的经过和盘托出，希望能够起到抛砖引玉的效果。”李部长洪亮的嗓门使嘈杂的教室即刻平静下来。

李部长刚一停顿，R钢铁公司的何总就按捺不住了，“李部长，国内工业界谁人不知，这五、六年，您跟徐教授偷经学艺，徐教授脑瓜的数据挖掘技术全移植到了你们企业。这几年，企业信息化建设和质量管理方面的国家级大奖几乎全被你们捧走了，你本人也升为教授级高工。谦虚什么呀，赶快讲吧，你们公司怎么开始与数据挖掘结上不解之缘的？怎么开展数据挖掘工作的？”

李部长把目光转向坐在最前排的何总，“急什么，何总，心急吃不了羊肉泡馍。昨天老孙家的羊肉泡馍刚一端上来，你就动筷子，烫着喉咙了吧。”逗得大家直笑……

2.1 李部长其人

李部长在T钢铁（集团）有限公司是个名人。

李部长叫李雪峰，1994年7月毕业于北京钢铁学院，到T钢铁公司当上了炼钢一车间的技术员。刚到企业不久，他发现公司生产的铸坯质量很不稳定，铸坯“夹杂”、“重皮”时有发生，公司老董事长甚为头疼。李雪峰主动向老技术员请教，从师傅们那儿，他发现了很多从书本上学不到的经验，真让他喜出望外。他把这些宝贵的一线操作经验总结归纳，编写成《转炉冶炼经验》，向工人传授。不久，炼钢一车间的铸坯质量明显高于其他两个车间。公司的老大难问题有了缓解，老董事长脸上露出了灿烂的笑容，举荐这个“初生牛犊”当了炼钢一车间主任。

新官上任，信心倍增，他并不满足这一点成绩，他深知公司的铸坯质量与国内同行还有较大差距，更无法与国外先进企业相比。下一步怎么办呢？大学四年，只学了些冶金学原理和生产工艺方面的课程，对冶金质量管理，一窍不通，真是书到用时方恨少！

他想到了母校，想到离校时带他毕业设计的导师孟教授曾嘱咐过“工作中遇到了什么问题，老师就是你的后盾。”于是，他来到了孟教授的办公室，滔滔不绝地详述了车间遇到的技术难题。他刚一讲完，孟教授就拿起钢笔，写下了四个字母“MSPC”，并风趣地说：“锦囊妙计，把我书架上的这本书带回去，好好研读。”



李雪峰如获至宝，当晚就踏上了回家的火车。一路上，他把《多变量控制》一书从头到尾看了两遍。天亮了，火车到了，他疲惫的脸上露出了希望的曙光。

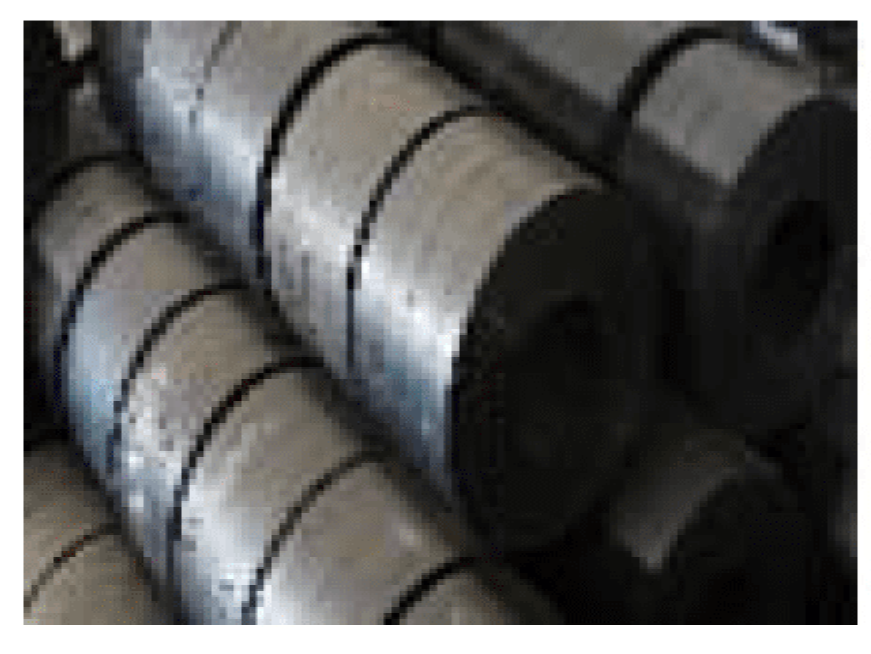
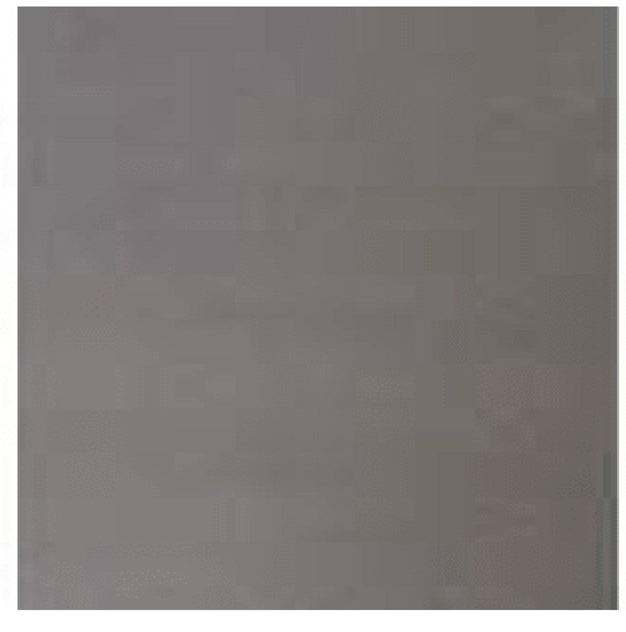
回公司后，他把自己关在办公室，奋战了三天，向老董事长提交了一份在本公司全面推广“多变量控制”的报告。他写道，“上世纪80年代以来，日本高质量产品的挑战使SPC（Statistical Process Control，SPC）在欧美工业界得到极大的重视。上世纪90年代初，统计过程控制被拓展为多变量控制（Multivariate Statistical Control，MSPC）。这种方法是应用主元分析（Principal Component Analysis，PCA）和部分最小二乘（Partial Least Square，PLS）等多元统计方法基于传统的统计过程控制而形成的一种对生产过程的多个变量进行监控、分析、控制的技术。MSPC应用的对象正是变量繁多的复杂生产过程的质量控制问题。建议公司尽快应用先进的MSPC技术提高产品质量，以使我们在激烈的国际竞争中立于不败之地。”

看了李雪峰的报告，老董事长激动不已，立即亲自主持召开了全公司中级以上技术人员大会，讨论通过了“应用MSPC技术，提高产品质量”的决议，并任命李雪峰为该项目的技术负责人。他边干边学，凭着良好的数学功底，很快掌握了MSPC的数学方法。通过MSPC技术，各个生产车间严格地对可能影响产品质量的人、机器、材料、方法和环境等因素进行全面监控，发现影响产品质量的不是工艺问题，而大多是工人操作不当、原料不达标、机器易耗部件不及时更新、不重视环境变化的影响等原因造成的。于是，李雪峰带着技术组的同事们，制定了细致的6σ管理策略，并建立了严格的生产操作规程。经过半年的努力，全公司的产品质量有了质的飞越，企业的经济效益大大提高。一年后，李雪峰被任命为公司品质部部长，他很快成了全公司的名人。十几年来，他任劳任怨，时时刻刻把握着公司每一种产品的质量脉搏。他多次放弃了提升的机会，他常说，质量是公司的命脉，品质部是公司的心脏，有了好的产品，企业才有出路。

2001年3月，李部长和公司其他几位同事考取了西安一所著名高校的工程硕士，2003年6月毕业。

2.2 老革命遇见了新问题

李部长打开了他的笔记本电脑，开始与大家分享他和他的同事们的数据挖掘之旅。他清了清嗓子，洪亮的晋西北口音使教室又恢复了平静：“话说2004年秋，我公司从德国引进了一套新的无取向硅钢生产线。2004年10月8日，是新硅钢生产线达产的日子，集团公司新上任的陈董事长和公司其他主要领导一大早就来到了生产车间。8点15分，一卷卷硅钢板缓缓下线。顿时，硅钢卷上折射出一道道闪光，现场响起了一阵热烈的掌声。公司领导个个神采奕奕，我更是格外高兴，真是谢天谢地，达产顺利。8点45分，领导们陆续离开车间，我也回到了办公室。”

突然，李部长眉头紧闭，右手狠狠地敲击了一下笔记本键盘，“世界上哪有这么容易的事。3天后，车间主任急匆匆地把我带到了卷取机前，让我查看了69卷有不同程度纵条纹的硅钢卷。”

李部长习惯性地用右手挠了挠头发，接着说：“当时我头一下子就懵了，硅钢卷中纵条纹严重而被打入废品的竟达31卷，另外38卷也因有不同程度的纵条纹不得不降级处理。估计因纵条纹缺陷每天直接损失不低于30万元。”从李部长的神情，学员们可以想象出他当时是何等的着急。

“不光我着急，硅钢生产线的所有工人和技术人员跟我一样。看着他们布满血丝的眼睛，我就知道，他们肯定这几天一直没有离开车间。”说到这里，李部长的眼睛湿润了。

“我让几个负责人留下，其他人立即回家，美美地给我睡上一觉。”李部长的话音刚落，电力公司的李总便开了口：“你手下的那些人，跟你是一路货色，爱厂如家，拼命三郎，肯定睡不着。”

说起他的得力部下，李部长更来劲了：“你说对了，他们一个也没走，钻进车间外面的汽车里临时休息，等候我的命令。”

急性子李总又发话了：“快说，你接下来有什么把戏可耍？”

李部长把视线移向了李总：“我顾不上他们了，先让几个技术人员说了说他们的看法。”

“有着近10年多变量统计质量控制经验的总工急不可耐，他说，当第8卷硅钢下线时，质检员就报告出现硅钢板表面出现纵条纹，我们没向领导汇报，自以为自己有十几年驾驭MSPC进行质量控制的经验，可经过3天的努力，就是发现不了引起纵条纹的元凶？”李部长沉重地说。

“6σ办公室主任也认为他们对每一个过程变量的控制也严格按照6σ管理规程操作，参数的命中率都足够高，可纵条纹仍然消失不了，真是奇怪透了！”李部长补充道。

“大家你一言，我一语，最后还是想不出有效的办法来。实在无奈，我只好将所有技术人员分为三组，分别跟班生产，密切注意纵条纹发展动向。第一组先留在这儿，其他人都回去休息。”李部长显出无奈的样子。

2.3　钓鱼钓来了数据挖掘思路

李部长沉默了一会儿后，轻轻地按了一下光笔，屏幕上出现了他钓鱼的照片。学员们都以为李部长按错了键，小声嘀咕起来。

李部长也看到了大家的诧异，急忙说道：“我的PPT没有放错，天无绝人之路，钓鱼钓来了数据挖掘，帮我们破解了硅钢纵条纹的技术难题。”



电力公司的王总疑惑不解，右手轻轻地敲了一下桌子："李部长，都到了燃眉之际，你还悠闲地钓鱼，鱼身上会有什么灵丹妙药？"

李部长走下讲台，看着电力公司的王总说道："且听我慢慢说来。转眼已经是周五了，我哪有心思回家。老婆打电话催几回了，说有几个老哥们在等我。我踏进家门，只见3个'鱼友'坐在沙发上。我明白了，他们想让我放松一下，换换脑子。"

清了清嗓子，李部长接着回忆："周六一大早，我们来到了南郊'渔乐园'。坐下来不到半个小时，伙计们个个捷报频传，可鱼儿就是不来造访我的浮漂。其实说真的，我虽然眼睛看着浮漂，但满脑子全是纵条纹。突然，浮漂动了，可我的电话铃也响了。是不是纵条纹问题有了进展，我急忙扔下渔竿，打开手机。原来电信公司短信息向我推荐'最近比较烦'、'曙光在前头'等彩铃。我气愤地合上了手机，起身想离开，但又怕打扰了伙伴们的兴致，只好又静坐下来。"

台下静悄悄一片，都在等李部长继续讲他的周末经历。

李部长长地吁了一口气说："我突然感到纳闷，为什么几个哥们和周围其他'渔友'没有收到这些彩铃推荐，电信公司偏偏对我情有独钟？为什么偏偏在自己心情不好的时候推荐了'最近比较烦'等类似的歌曲？傍晚回家时'鱼友'们个个满载而归，一路上兴致勃勃，而我却一言不发，一连串的疑问让我百思不得其解。回到家中，我便向在大学电信学院教书的同学打了个电话，请教其中的奥秘。"

台下的张行长心急火燎地问："李部长，你同学给你揭秘啥？"

李部长笑盈盈地说："他告诉我，电信公司对用户的信息进行了数据挖掘，并向我解释了其中的门道。原来如此！钓鱼前的某天晚上，我在微博上将最近生产中遇到的问题简单描述了一下，期望有同行帮忙，并写了'郁闷'、'着急'之类的话语，而且以前我也咨询过彩铃业务，于是电信公司就把我作为潜在客户进行精确营销。"

看着李部长夸电信的主动营销做得好，冯总甭提脸上多有光彩了。

李部长思路严谨，接着讲述："我当时一个激灵，想起来在工程硕士班上《最优化及其应用》课时，老师曾提起过数据挖掘在工业生产中的应用。此时我不知从哪儿来了一股劲，非常渴望了解一下数据挖掘技术。"

受李部长感染，黄主任说："李部长，看你对知识的渴望劲儿，是不是预感数据挖掘可能就是解决燃眉之急的良方了？"

李部长肯定地回答道："可不是么，我急切地在百度上输入了关键词'数据挖掘'。我发现数据挖掘在国内外都是研究的热点，而且在互联网、金融、电信、商业、交通、电力、政府机关、工业生产等领域都有很多成功的应用案例。当'数据挖掘在钢铁产品质量控制中的应用'这几字映入眼帘时，仿佛抓住了一根救命稻草。我急切地浏览着这方面的内容，了解到数据挖掘技术在冶金行业特别是钢铁生产中已经有不少成功的应用。例如，安阳钢铁公司在板坯连铸的二冷配水中应用数据挖掘技术，解决钢板裂纹问题；宝钢在钢材产品质量管理、配矿优化、节约运输成本等方面的成功应用；湖南冶金总公司将数据挖掘技术应用到焦化配煤优化中，不仅使焦炭质量提高，也大大降低了生产成本。"

姚局长说："李部长，我能理解您当时的激动心情，这些成功案例可都是和您面

临的问题息息相关呀！这下你心中的信心之火被燃起来了吧？”

李部长感慨地说：“确实是，搜集到的信息越多，我才知道自己平时忙于生产管理，与研究单位和高校接触太少，被新技术远远地抛在了后面。我再也按捺不住内心的激动，觉得数据挖掘必定可以解决硅钢纵条纹质量问题。于是我连夜向公司领导写了一份《应用数据挖掘技术解决硅钢纵条纹质量控制问题》的报告。”

2.4　数据挖掘项目立项

李部长回忆起了次日的情形：“第二天早上8点，我带着这个报告来到陈董事长的办公室，只见集团公司总经理、总工都在。董事长看我来了，风趣地说：‘说曹操到曹操便到’！刚才我与两位老总商量好了，准备交给你一项开创性的任务。”



“我心想，肯定是硅钢纵条纹问题要给我下死命令了。幸亏昨天晚上已有准备，我急忙将写好的请命书递了上去。三位老总看了报告的题目，都笑了。我不知道他们在笑什么。”

董事长当时看着我不解的样子，解释道：“异工同曲，我与业界同行和有关专家沟通过了，硅钢纵条纹问题可以尝试应用数据挖掘技术解决。我给二位老总建议让你啃这块硬骨头呢，刚准备给你打电话你就自报家门来了。我们集团公司这位MSPC的开拓者，又将成为数据挖掘的先行者了。”

移动公司梁总说：“真是英雄所见略同啊，李部长，您和领导想一块儿去了！”

李部长有点惭愧地说：“说实在的，数据挖掘能不能消除硅钢纵条纹缺陷，我心里一点也没底。但我预感到，即使不能彻底解决问题，起码会有一定的效果。于是我干脆地答道：“请老总们放心，我们尽力完成任务！”

李部长说完最后一句话的时候声音特别洪亮，学员们一阵阵鼓掌，为他加油喝彩。

掌声刚落，电力公司王总就开口了：“李部长，这回你可是‘滚油锅里捡金子——无法下手啦！’”

李部长目光移向电力公司王总：“吉人自有天助！董事长也给我了一张神秘的小纸条。”

王总急了：“是不是消除硅钢纵条纹缺陷的灵丹妙药？”

李部长右手轻轻在键盘上一敲，屏幕上出来了一个人的头像。“是这个人的联系方式。”

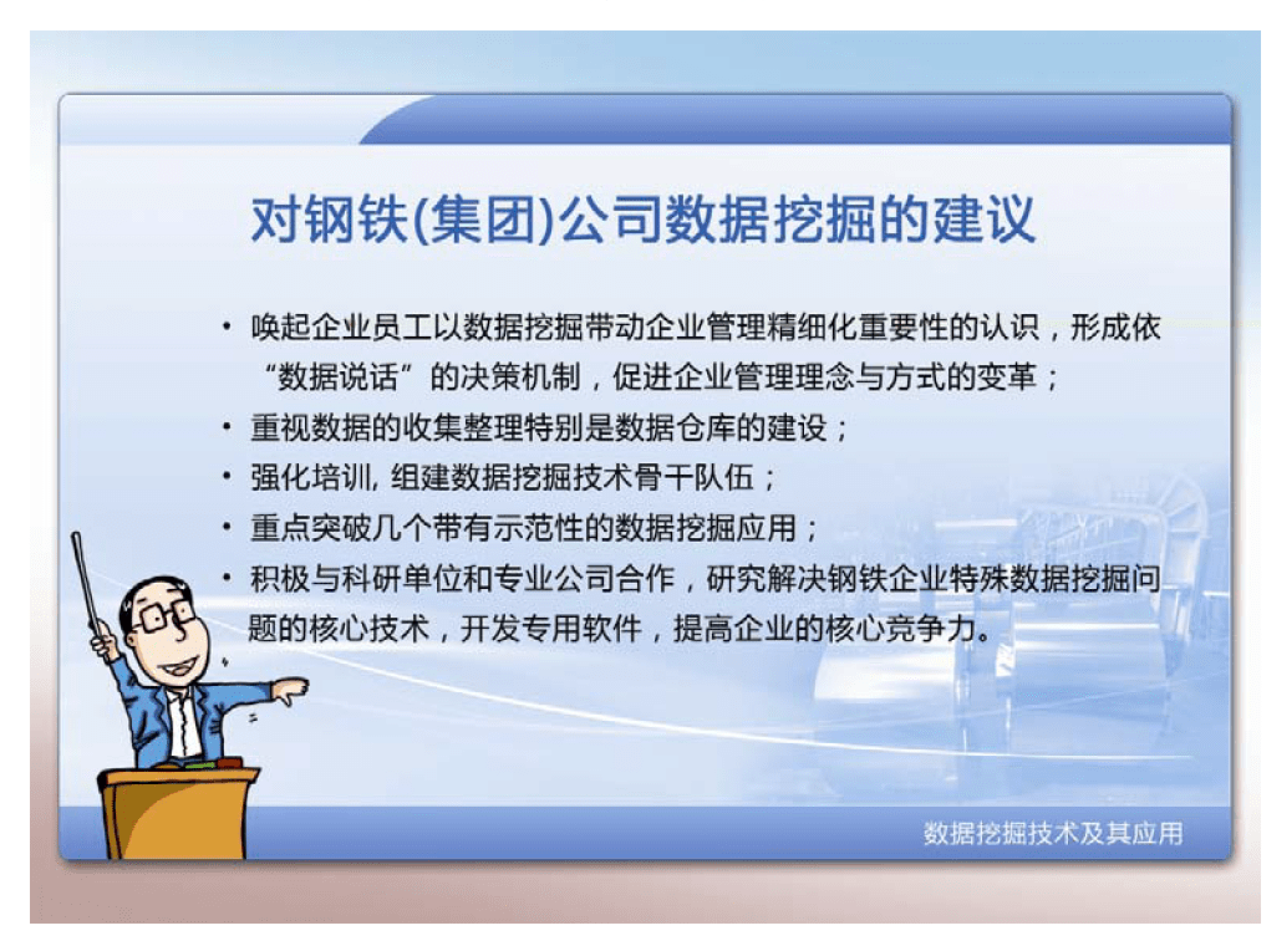
学员们一看，都笑了，齐声喊道：“徐教授！”

李部长喜形于色：“原来，董事长读研时，听过徐教授的《智能计算》课，董事长早就与徐教授探讨过应用数据挖掘技术进行流程工业质量控制的方法。”

听李部长这么一说，大家也都被徐教授的影响力折服了！

李部长接着刚才的话题补充说："回到办公室，我拨通了徐教授的电话，描述了我们生产中遇到的技术难题，并向徐教授求援，他欣然答应。三天以后，徐教授带领六人教授团来到了公司，他们个个都是智能信息处理的专家。集团公司领导陪同专家们参观了硅钢生产线后，双方进行了深入的交流。随后，徐教授向公司技术人员作了《数据挖掘技术及其应用》的报告，向我们讲述了数据挖掘基本概念、典型任务、核心技术，并对我们公司开展数据挖掘工作提出了一些建议。"

李部长轻轻按了一下光笔，屏幕出现了如下内容：



李部长将光笔指向建议的第一条，解释说："钢铁企业是流程化的生产单位，虽然生产自动化程度非常高，但是，老实说，我们的很多工序（如炼铁、炼钢、连铸、轧钢等）的过程控制很大程度上依赖技术工人的经验，对生产过程的驾驭还比较粗放。不过，我们已经建立了先进的信息化平台。尤其是近几年企业形成的'建设创新型企业'的文化氛围下，我们公司积累了丰富的数据，也具备了一支高素质的管理技术队

伍。企业高层领导一致认为，科学决策是企业信息化建设的最终目标，数据挖掘是实现这一目标的有效工具，是构筑未来核心竞争优势、保持可持续发展、实施精细化管理的战略选择。”

李部长刚一停顿，国内产能最大的 S 钢铁公司的赵总起身问道：“李部长，据我所知，我们两家公司一样，都投资数亿元引进世界 500 强 SAP 软件公司的钢铁生产管理解决方案，它是一个全面、完整、集成的系统，其功能覆盖了财务、成本、生产、销售、供应、库存、质量、项目管理、设备维护、人力资源管理、供应链管理、客户关系管理、供应商关系管理、决策支持等钢铁企业信息化管理各方面的需求，可见该系统的数据可谓包罗万象、应有尽有，可徐教授为什么还要建议重视数据的收集整理，特别是数据仓库的建设呢？。”

听了赵总的问题，李部长笑了：“你的疑虑跟我是一样的，我也问过徐教授这个问题。大家知道，SAP 系统其实就是 ERP 系统，它以供应链为主线，包括从销售订单或生产经营计划→生产排程→组织采购→安排生产→销售发货的整个过程，着力于计划流、物流、信息流、资金流的统一运转，通过计划流驱动物流，通过物流驱动资金流的良性循环。从 ERP 的角度来看，SAP 系统确实不辱‘全球最佳’这一称号。但从数据挖掘的角度着眼，我们需要关注新产品设计、改进产品质量、降低生产成本、设备故障检测等这些主题。这些方面涉及到基础自动化（L1）、过程自动化（L2）、产线管控（MES）、经营管理（ERP）、决策支持（DSS）等信息系统。可是这五级系统并没有完全整合，在一定程度上还是‘信息孤岛’。当确定了数据挖掘的目标后，就需要对数据进行整理。当然，像我们这样正在进军世界 500 强的大型钢铁公司，可以通过数据挖掘解决的问题太多了，最好是统一规划，建立数据仓库。”

赵总边听边点头：“信息孤岛，害人不浅！有一次我们要分析钢材表面夹杂缺陷的原因，各车间的生产数据在各自的生产系统中，而且数据缺失、噪音比较严重，技术人员花了十天左右时间对相关的数据进行清理、整合。”

赵总端起水杯，刚准备喝水，又放下杯子，问道：“李部长，数据挖掘项目与一般的信息化项目一样，主要由专业公司或科研单位来完成，钢铁公司相关人员配合就

行，徐教授为什么还建议‘加强培训，组建数据挖掘技术骨干队伍？’”

李部长急忙道：“赵总，对这个问题我原来与你的认识一致，现在我体会更深了，数据挖掘项目与普通的信息化项目还是有很大差别的。目前我国的大中型企业不乏信息化方面的技术人员，但懂得数据挖掘的人才寥寥无几，在这种条件下开展数据挖掘工作，一方面需要与高校等科研单位或专业的数据挖掘公司合作，另一方面还要加强数据挖掘知识培训，培养一些既精通本领域业务，又熟悉数据挖掘流程，了解数据挖掘方法的技术骨干。这样，行业领域技术人员和数据挖掘专家一起才能从实际工作中提炼出可以通过数据挖掘方法解决的问题，建立合理的数据模型，客观地评估数据挖掘的结果。”

汽轮机公司的江总听到这里，打断了李部长：“不就是需要人吗，现在公司里硕士博士一大群，一个个好学上进，与数据挖掘专家一起组建一个开发组不就行了！”

看着江总咄咄逼人的样子，李部长有点不服气了：“江总，你以为只有你们公司人才济济。现在，连 2、3 百人的小企业都有几十名硕士，不信，你问问在你前排就座的玻璃公司的彭总。”

彭总会意地点了点头。

李部长接着说：“组建了团队以后，怎样开展工作呢？大家首先要清楚地认识到，数据挖掘可以解决企业生产、管理中的很多用常规方法难以处理的问题，但数据挖掘也不是万能的，不能包揽所有问题。而且还会有一些问题应用经典的数据挖掘方法无法得到满意的结果，需要数据挖掘专家针对具体问题建立相应的数学模型并设计特有的求解算法才能解决。因此，开展数据挖掘的初期，最好选择一些相对容易的问题，这样，一方面能够很快领略到数据挖掘的奥妙，另一方面为解决较为复杂的问题积累经验。”

汽轮机公司的江总又开口了：“我明白了为什么徐教授专门强调‘重点突破几个带有示范性作用的数据挖掘应用’。对头，旗开得胜，往后不要命！后来你们选择了几个问题试图应用数据挖掘方法解决？”

李部长回答："经过与徐教授等人反复讨论分析，我们认为硅钢纵条纹问题是我们迫在眉睫、不能回避的问题。虽然有相当的难度，但也得背水一战。在硅钢纵条纹项目完成后，我们继续进行基于支撑向量机和遗传算法的热连轧质量控制方法研究。经公司领导同意后，我们钢铁公司和数据挖掘公司先签订了消除硅钢钢板纵条纹缺陷的数据挖掘方法研究技术协议。双方决定共同组建数据挖掘团队，团队由专家组、数据组、算法组、软件组和部署组 5 个组构成，由李部长担任甲方数据挖掘项目经理，负责整体负责数据项目的实施。由数据挖掘公司的卢经理担任乙方项目经理，具体开展数据挖掘工作。"



2.5 数据挖掘项目实施

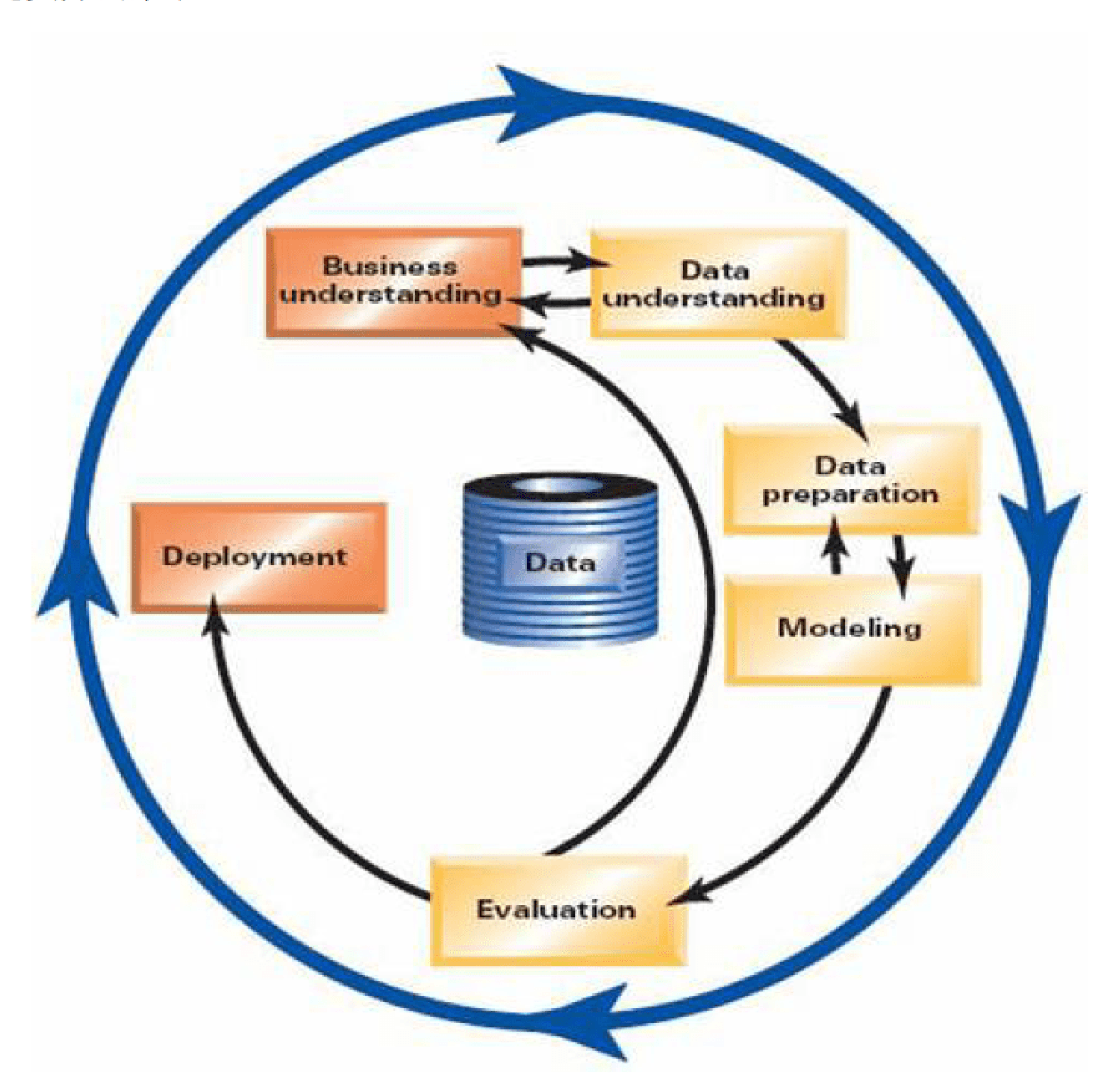
"李部长，这回你可谓骑马上独木桥——回不得头了！" S 钢铁公司的赵总笑嘻

嘻地说。

李部长显得不慌不忙："有了徐教授的数据挖掘研究中心作坚强后盾，我信心十足。研究团队成立后各小组立即紧锣密鼓地按照'跨行业数据挖掘标准流程'既有分工又相互协作地开展工作，经过一个半月的奋战，终于取得了可喜的成果。"说到这里，李部长脸上露出了灿烂的笑容。

这时，汽轮机公司的江总却眉头紧皱，不解地问："李部长，'跨行业数据挖掘标准流程'是不是就是一种通用的数据挖掘过程，你还是给我们介绍介绍吧！"

李部长说："不好意思，咱不是专业教师，犯了这样低级的错误，将没有讲解的名词先卖弄出来了。"



李部长用光笔指着这张流程图说："为了低成本、易操作、高效、可靠地进行数据挖掘，经过数据挖掘标准化联盟对十几年数据挖掘实践进行经验总结和理论抽象，创建了跨行业数据挖掘标准流程，即 Cross Industry Standard Process for Data Mining，简称 CRISP-DM。它包括业务理解、数据理解以及收集、数据准备、建立模型、模型评估和部署六个阶段。我们消除硅钢钢板纵条纹缺陷的数据挖掘方法项目也是按照这六个步骤进行的，下课铃响了，休息一会我再给大家较为详细地说明。"

2.5.1 业务理解阶段（Business Understanding）

李部长从来没有讲过这么长时间的课，有点累了。他来到教授休息室，工作人员马上给他递了一杯热腾腾的牛奶，他一饮而尽，又要了一杯。

上课铃响了，李部长精神抖擞地走进教室，手里还端着未喝完的牛奶。

汽轮机公司的江总看着李部长的样子，开玩笑道："李部长，你光顾自己享受教授级待遇了，也不给你老哥讨杯牛奶来，别忘了，刚才那节课咱们俩一问一答，我都成'助教'了。"

"这不，还有半杯，你喝吧！"李部长将杯子递向江总，江总急忙摆了摆手。

李部长走上了讲台："不开玩笑了，咱们接着上课。江助教，上一节课讲到到什么地方了？"李部长装出一副若有所思的样子。

"跨行业数据挖掘标准流程。"江总喊道。

"哈哈！其实我是想转移你的注意力到课堂上来。"李部长边说边打开笔记本电脑。

这时，屏幕上跳出了如下内容：



李部长用手中的光笔指着投影："我代表甲方提出，硅钢纵条纹问题的需求很明确，就是要应用数据挖掘方法找出导致纵条纹问题的关键因素，并实现对关键因素的控制达到消除硅钢纵条纹的目的。"

S 钢铁公司的赵总："李部长，你们的要求太宽泛了吧？虽然我是数据挖掘的外行，但起码明白不管干什么事情，目标必须非常具体，在产品质量控制方面更应当如此。"

李部长知道，赵总在 S 钢铁公司主要负责产品质量管理，指挥过无数次质量问题技术攻关，他的话真是一针见血。

于是，他将目光转向赵总："是的，赵总说得太对了！在第一次数据挖掘会议上，我先汇报了硅钢生产线出现纵条纹缺陷的情况。我们公司技术中心教授级高工刘主任从冶金学原理方面陈述了纵条纹产生的机理，轧钢厂杨总工描述了硅钢生产流程并分析了影响硅钢纵条纹的因素。X 大学数据挖掘中心金教授介绍了对硅钢纵条纹问题数

据建模的初步设想。经过各小组成员一起认真分析认为，硅钢纵条纹问题有望通过非平衡的分类方法解决。最后，提出了将硅钢纵条纹比率由现在的 12.1%降低到 1.8%的目标。”

听到这里，赵总激动了：“要降低 10.3 个百分点，难啊！”

李部长倒是胸有成竹的样子：“事在人为嘛，只要努力就有成功的希望。不过，不能光吹牛，关键还在行动。我们制定了详细的数据挖掘计划，要求各组分工协作，紧密配合，争取在两个月内完成任务。”

“两个月攻克硅钢纵条纹难题，目标定那么高，时间又如此短，完不成任务，看你咋给董事长交待！”S 钢铁公司的赵总替李部长捏一把汗，喃喃道。

李部长不慌不忙地说：“我们有目标、有计划，有一支由冶金专家和数据挖掘专家组成的攻坚团队，更重要的是还有董事长的大力支持，万事俱备，我们只需要按照数据挖掘的流程一步一步坚定地走下去。”

2.5.2 数据理解阶段（Data Understanding）

李部长抬起右手，使劲地敲击了一下笔记本电脑的回车键，大声说：“Go！下一步我们进入了数据挖掘的数据理解阶段，请看大屏幕。”



“在这一阶段，我们根据硅钢纵条纹产生的机理和硅钢生产流程，经过反复筛选，初步确定硅钢纵条纹的影响因素有连铸中包温度、连铸拉速、铸坯成分、粗轧出口温度、精轧出口温度和卷取温度等共 21 个。”李部长如数家珍地说。

S 钢铁公司的赵总又开了口：“这些数据分布于转炉冶炼、连铸、加热炉加热、热轧粗轧、热轧精轧、常化/酸洗、退火和剪切等工序。据我所知，你们公司还未建立数据仓库，数据需要从相应部门的数据库中提取，这些部门可不一定神速地执行你李部长的指令！”



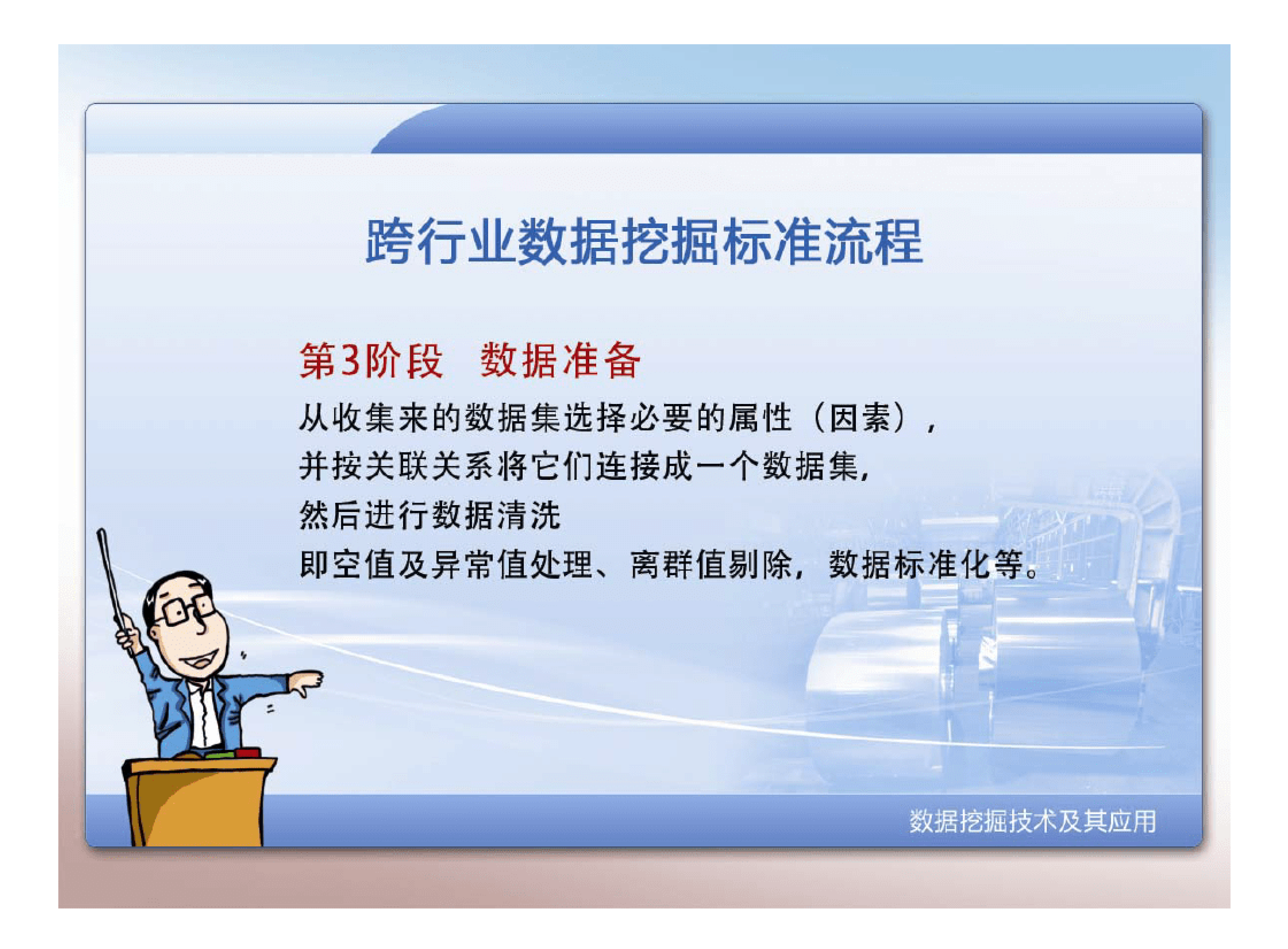
李部长得意地说："我有董事长的尚方宝剑，底下哪些小头目们岂敢怠慢。我们只用了 5 天时间，数据组就将数据从相关部门收集来了。他们浏览各部门的数据，发现数据有不少缺失，甚至还有明显的异常。进一步分析发现，有些影响因素的数据方差特别小，于是便将它们认为是常量。数据组一致认为虽然从理论上说这些因素对硅钢纵条纹有作用，但生产工艺控制命中率足够高，使得相应的影响因素数据变化很小，对硅钢纵条纹的作用几乎恒定不变。于是将这些影响因素删除，影响因素从原来的 21 个减少到 15 个。最后，数据组给出了影响纵条纹的因素列表，并对数据具体含义、命中目标值、异常、缺失等进行了详细的描述，形成了《数据收集及质量检验报告》。"



2.5.3　数据准备阶段（Data Preparation）

李部长一口气讲了这么多，端起水杯咕噜噜地喝了起来，S 钢铁公司的赵总趁机开了口：“下面该到数据挖掘的第 3 阶段了吧？”

这时李部长赶紧敲了一下键盘，屏幕出现：



李部长解释说：“数据理解阶段已经初步确定，硅钢纵条纹的主要影响因素有 15 个，包含连铸中包温度 t1、t2、t3，连铸拉速 v1、v2、v3（数据来源于连铸数据库），铸坯成分 C、Si、Mn、S、P、Al（数据由检化验数据库获得），粗轧出口温度 RT0、精轧出口温度 FT6 和卷取温度 CT（要从轧钢数据库提取）。这些数据可由铸坯编号、转炉编号和硅钢卷号关联形成一个数据表。然后再对这个表进行空值

及异常值处理、离群值剔除操作。”

听到这儿，赵总站了起来，大声吼道：“李部长，手下留情啊！我们 S 钢铁公司开展数据挖掘几年了，数据仓库都建了，信息中心汇报数据准备情况时从来没提过剔除数据！”

李部长：“我非常理解赵总的心情，一般最好不要轻易删除数据，对于空值、异常值处理、离群值通常采取均值、迭代回归等方法进行补缺或修正处理，尤其在样本数量较少的情况下更应当如此。不过经过 1 个多月的生产数据积累，我们采集的数据量比较充分，删除极少量‘坏’样本对数据建模不会有什么影响。”

赵总指了指屏幕：“李部长，数据清洗好了，PPT 上为什么说还要进行数据标准化？”

李部长笑道：“这个问题问得很好，起初我也不知晓其中的道理，听了 X 大学 Merit 数据挖掘中心金教授的解释我才明白了其中的道理。对了我笔记本电脑上保存着金教授对我公司进行数据挖掘培训的录像，咱们一起欣赏一下数据标准化这一段吧！”

视频播放器刚一关闭，赵总就侃侃而谈自己的心得体会：“我明白了，是有这样的问题——采集的数据数量级上相差太大了，我知道铸坯成分 Al、Si 为百分之零点几，C、S、P 为百分之零点零几，而粗轧出口温度 RT0、精轧出口温度 FT6 和卷取温度 CT 均高达好几百度，如果不进行数据标准化，计算时可能出现大数吃掉小数现象，导致得到的模型误差太大。”

李部长觉得赵总理解得还真够到位，又问了一句：“赵总，那一般用什么方法对数据进行标准化处理？”

赵总挠了挠头：“刚才金教授好像提到常用的数据标准化方法有……，有好像‘0 均值–1 方差法’、‘最大值–最小值法’和‘移动小数点法’等。”

李部长拍手称道：“别看咱赵总都快奔 5 的人了，记性还不错。数据标准化通常多采用第一种方法，即将变量数据化为‘均值为 0，方差为 1’范围内的数据。”

赵总更加得意了："我的记忆力可好了，我还记得我们公司负责数据挖掘的孟博士说过，数据预处理阶段太重要了，这一阶段的工作是保证整个数据挖掘成功的关键。"

李部长打了个暂停的手势："说你胖你就哼。"惹得大家直乐。

2.5.4　建模阶段（Modeling）

这时李部长表情有点凝重，他的 PPT 翻开了新的一页：

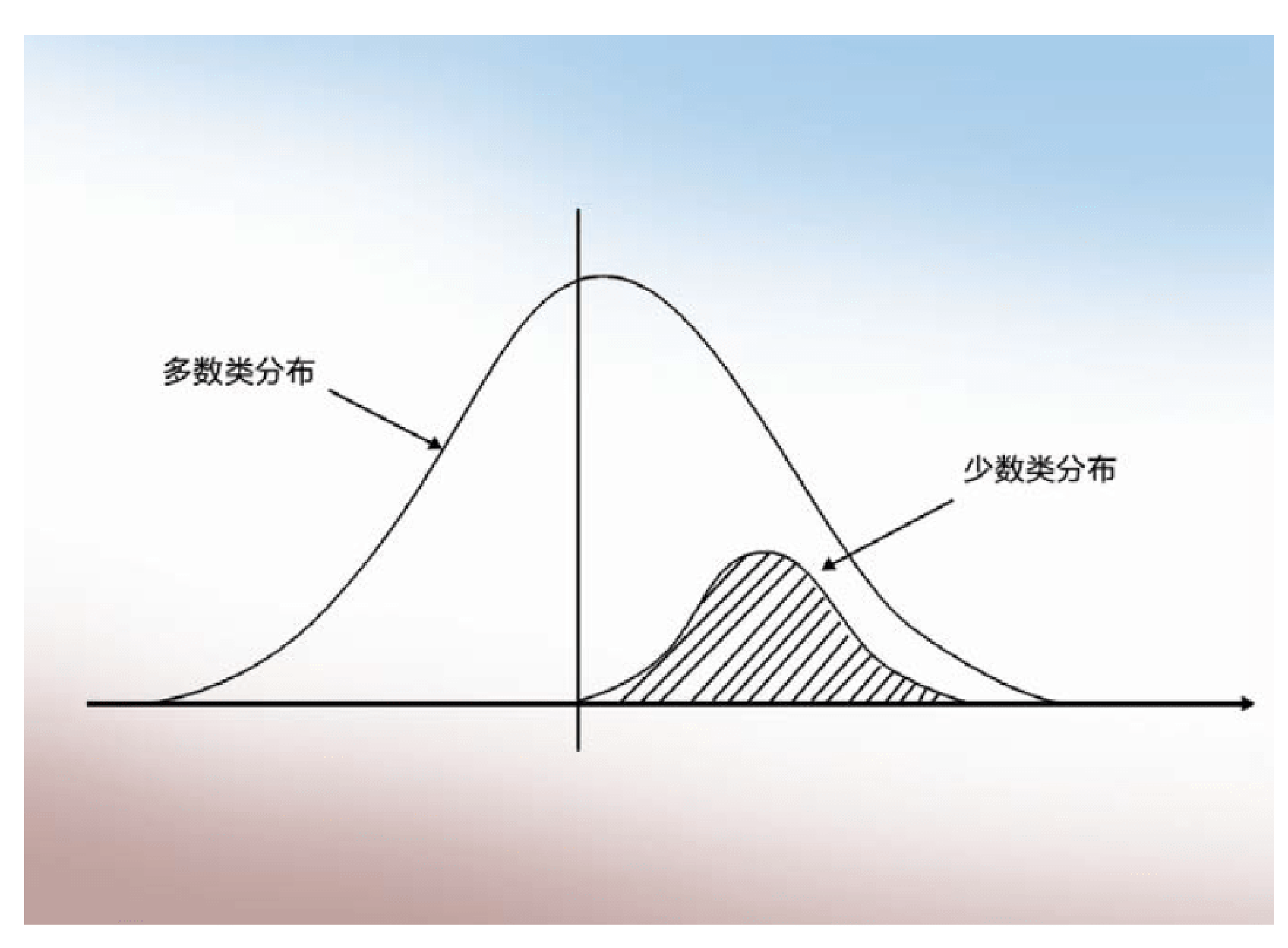


他指着大屏幕说道："数据挖掘流程的第 4 阶段的数据建模主要由 X 大学 Merit 数据挖掘中心完成。中心的金教授说，硅钢纵条纹问题初步分析就是一个非平衡分类问题，可他们将几乎所有的分类问题的数学模型和求解算法统统试验了多遍，所得到

模型的预测能力都非常差。后来徐教授亲自坐镇研讨了数次，发现硅钢纵条纹数据集不仅是非平衡数据集，而且是不相容数据。”

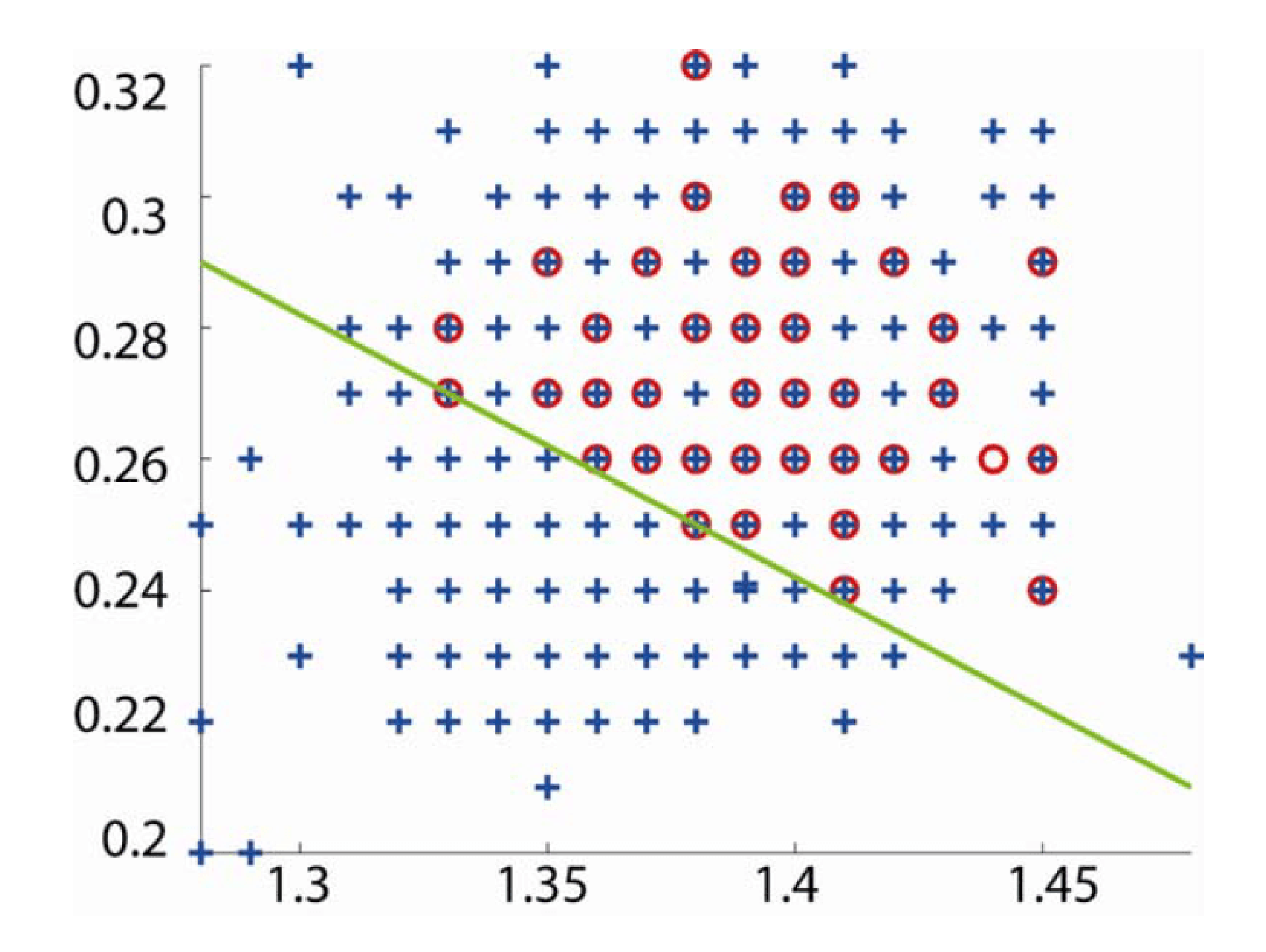
赵总有点诧异：“李部长，非平衡数据集你刚才提过，现在又冒出个不相容数据，这到底是什么意思呢？”

李部长侃侃而谈：“硅钢生产是非常复杂的生产过程，产生纵条纹的影响因素很多，为了简化问题和方便数据获取，我们忽略了一些对纵条纹作用相对较小的影响因素，这样就会存在很多硅钢产品，其影响纵条纹的因素非常相同或相近，但纵条纹的类别完全相反。这样的样本称为不相容样本，相应的数据集称为不相容数据集。”



赵总从李部长手上拿过光笔，指向图上的红点：“李部长，这些红点大部分中还套有蓝色的‘+’号，是不是这些样本就是不相容数据？”

李部长点了点头，连声说：“对，对。”



S 钢铁公司的赵总对这张数据示意图极为感兴趣："我是 X 大学计算机系数据挖掘专业研究生毕业，从事数据挖掘工作也有五、六年了，据我了解，对非平衡数据分类问题的研究近十年来一直是国内外很多学者关注的热点，而不相容数据建模问题却少有人研究。李部长，请您介绍一下对硅钢纵条纹问题的数据建模方法吧。"

李部长有些为难的样子："这个我可说不好，不过大概思想我还是清楚的。"他向赵总要回光笔，将光点指向图的左下方，继续说道："不知大家留心没有，图的左下方全是蓝色的'+'号，代表这一片区域都是正品，是生产的'优区'，右上方蓝色和红色交叠，表明这部分区域次品正品都有，是生产的'劣区'。我们只要'使生产在优区进行'的规则就行了。"

尚主任眉飞色舞，激动地拍了一下桌子："有道理！快讲一下具体是如何建模的。"

李部长笑道："你就别赶鸭子上架了。两年前，金教授曾经代表 X 大学 Merit 数据挖掘研究中心详细地介绍了对硅钢纵条纹问题建立的数学模型和设计的求解算法，我只记得叫作 L1 正则化模型……，我找一下金教授当时的 PPT。"

找到了以后，李部长接着说："对了，他们先提出了一种新的分类准则，称为支

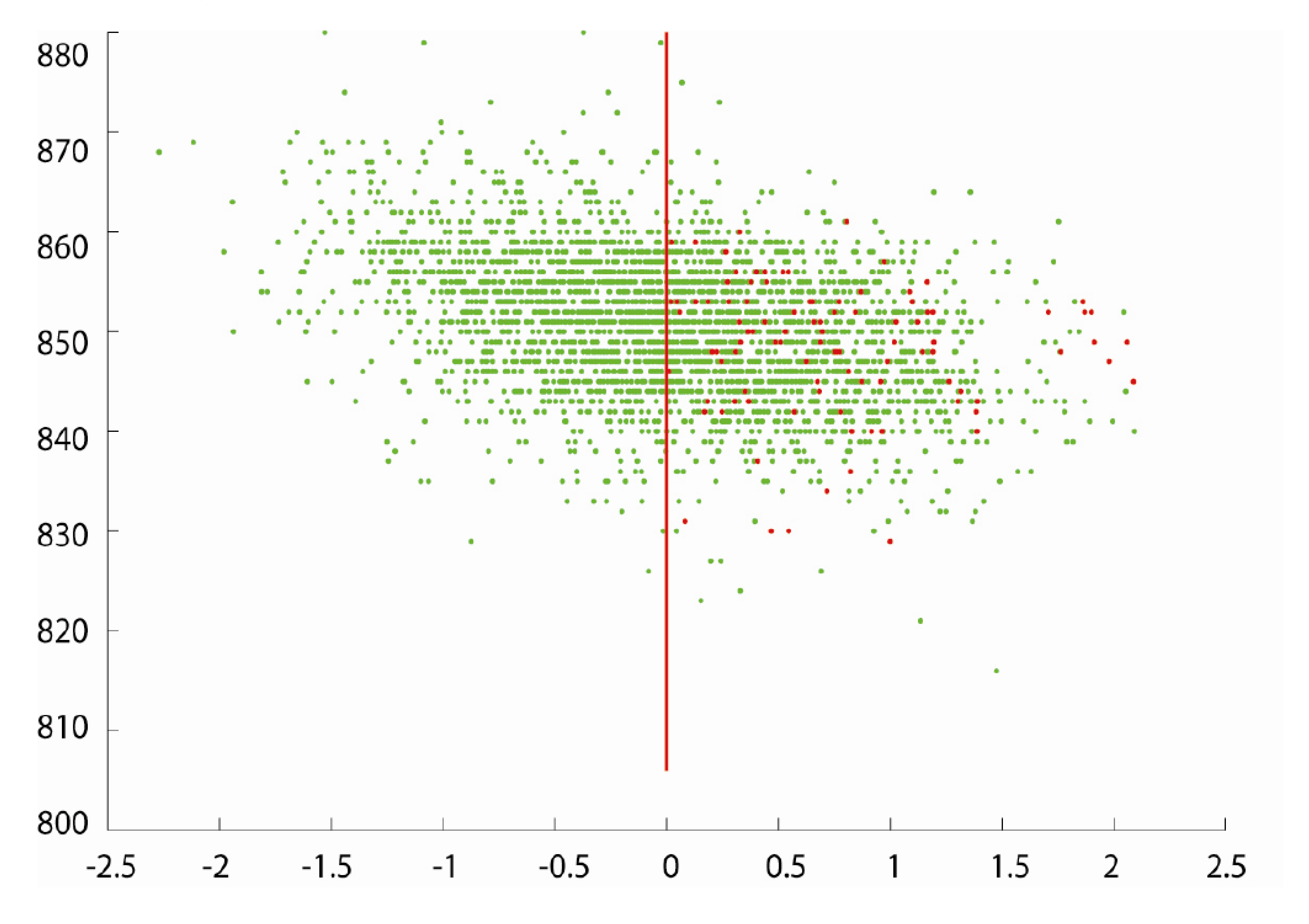
持度最大化准则，即分类器分出的‘优区’的样本尽可能的多。还提出了实现支持度最大化准则的代价敏感损失函数，在此基础上才建立了消除硅钢纵条纹缺陷的 L1 正则化模型，还创新性地设计了求解该模型的 L1 稀疏迭代算法。”

尚主任眉头越来越紧：“且慢，且慢！我越来越听不懂了，您还是说详细点吧。”

李部长干脆拿起粉笔，在黑板上写了一个电子邮箱的地址：wfjin@hotmail.com ，说道：“我也是只知道几个名词而已，大家谁对这个方法感兴趣的话，给金教授发电子邮件联系。他这个人可热情了，有问必答。”

尚主任高兴地叫道：“那好，这个周末我就去拜见金教授。”

李部长将 PPT 又翻了一页，屏幕上出现了一个图，他指着这个图说：“你去的时候最好带着这张图，金教授看见这个图就会高兴得很。”



尚主任不解地问：“难道这是一张联络图？”

李部长笑道："这是硅钢纵条纹问题模型求解结果的展示图，谁看了它都赞叹不已。"

尚主任更感兴趣了："那您就给大家解释一下这张图表示的意义吧！"

李部长的光笔指向图中间的直线："这条线就是模型求解得出的生产'优区'和'劣区'的分界线，线的左面全是正品，优区样本的数量对所有样本的比率（即支持度）高达 49.11%。如果将生产控制在优区进行，就会极大降低硅钢纵条纹出现的几率。"

尚主任激动不已，连声赞道："原来如此，妙，妙！"

2.5.5　模型评估阶段（Evaluation）

李部长右手一挥，使劲地敲击了一下键盘，并说："妙什么呀！还没有进行模型评估呢！请看屏幕。"



李部长说："模型评估是至关重要的一个环节，未经过评估的模型千万不可直接就去应用。因为所得出的模型只是通过已有的数据得出，对未来数据的预测能力如何，一定要经过实践的检验。"

S 钢铁公司的赵总顺口便问："那你们是如何检验所得到的'优区'和'劣区'分界线的可信度？"

李部长兴高采烈地说："真是功夫不负有心人呐，后来半个月生产所出现的纵条纹样本全部落到了'劣区'。"

沉默好久的尚主任又问了一个关键性的问题："所得到的'优区'和'劣区'分界线的可解释性怎么样？"

李部长不慌不忙地说："我们获得的分界线（即分类器）是线性的，我们可以根据每一个变量前面的系数的正负判断其对纵条纹是正面影响还是负面影响，依据其绝对值的相对大小衡量相应的变量对纵条纹的作用大小。从分类器的表达式容易看出，Si、FT6、Al 和 P 为硅钢纵条纹的主要影响因素，这与理论分析的定性结论相符。"

尚主任眼睛一亮，大声叫道："那么，下一步就可以放心地跨入数据挖掘的部署阶段了！"

2.5.6 部署阶段（Deployment）

李部长点了一下鼠标，喊道："下一步，部署阶段。"

屏幕上出现了如下画面：



李部长：“我们将原来生产控制策略中影响硅钢纵条纹的 15 个因素的命中目标值代入所得到的分类器中，发现它正好位于‘优区’和‘劣区’分界线偏右处。可见，这正是硅钢纵条纹比率高的原因。为了保持生产的稳定进行，我们只对硅钢纵条纹影响最大的 4 个因素的命中目标值作了调整，将调整后的 15 个影响因素的目标值代入所得的分类器中，结果落入‘优区’和‘劣区’分界线的左侧。”

尚主任：“这么说新的生产控制策略是可行的？”

李部长：“我们将一个半月来的数据挖掘工作进行了详细总结，最后完成了《应用部署报告》，上报公司领导批准实施改进的生产控制策略。”

“领导反映怎么样？”尚主任急切地问。

李部长铿锵有力地回答道:“董事长召集公司技术中心硅钢研究室的几位研究员、硅钢生产线的主要技术人员和国内著名硅钢专家 W 钢铁公司的施总工对我们改进的控制策略进行了反复论证，最后同意了我们的方案。”

“实施结果怎么样？”尚主任迫不及待地追问。

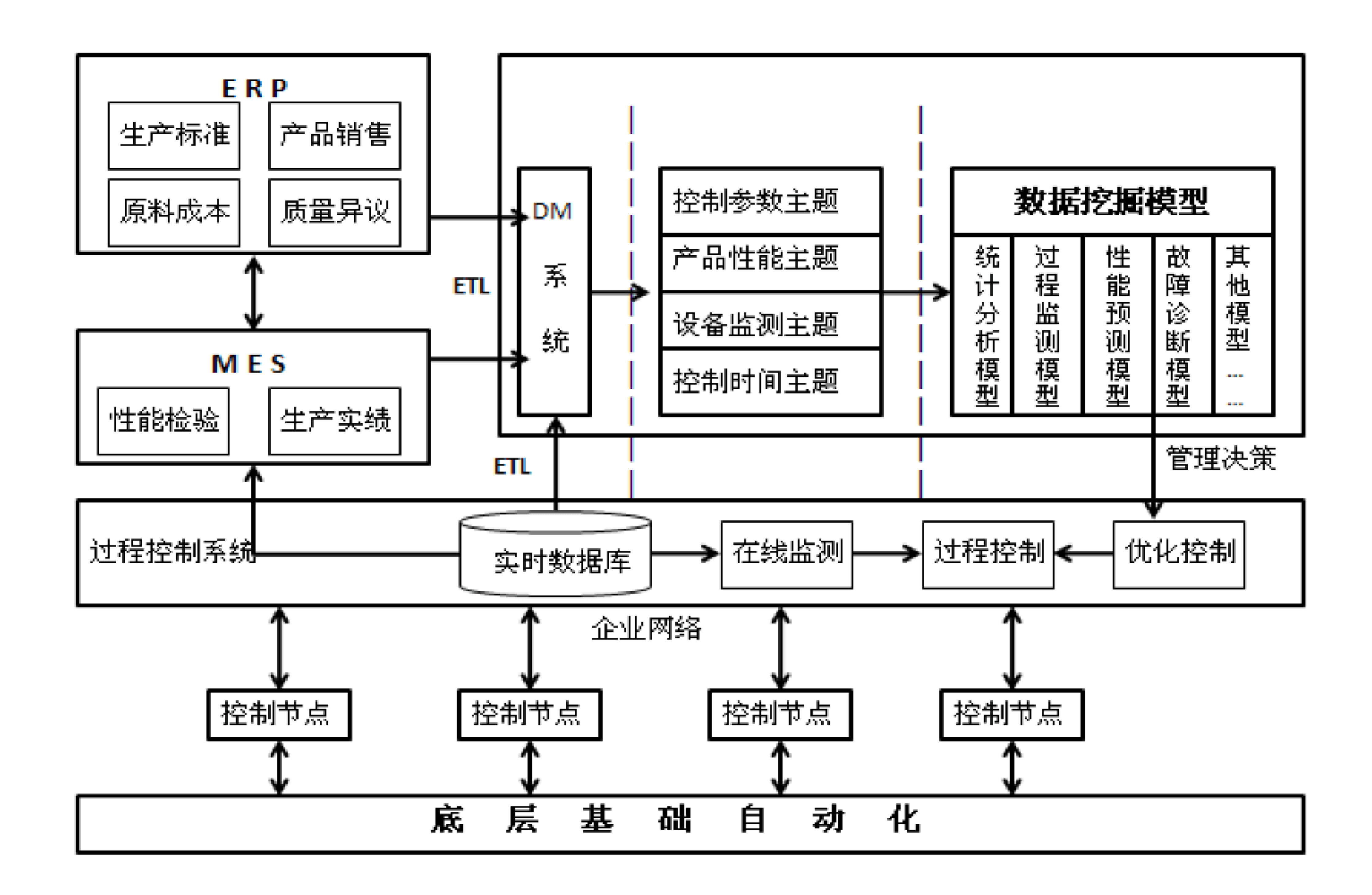
“功夫不负有心人，我们成功了，新方法效果明显。一个月后统计结果令人振奋，硅钢纵条纹的比率降低到了 1.65%，产品的各项性能指标达到了国际先进水平。真是‘靓女不愁嫁’，三个月后，我公司各种牌号的硅钢在国内外市场成了抢手货。”李部长越说越激动。

李部长话音刚落，教室里便响起了一阵热烈的掌声。

2.6 李部长的展望

经过纵条纹数据挖掘项目的实施，李部长再一次成为公司的名人，但他并没有沉溺于成功的喜悦中，他思考着如何利用数据挖掘做更大的事，为公司谋取更大的利益。

李部长和数据挖掘公司的卢经理促膝长谈了一次，觉得数据挖掘技术在钢铁行业的应用不仅限于质量控制。企业建立起的生产过程实时数据库、ERP 系统等每天数据量以 3Gb 增加，数据越来越丰富，应尽快建立数据仓库。经过一翻交流后，卢总给李部长拿出了企业级数据挖掘系统规划方案：



利用数据挖掘技术实时建模，可以快速实现企业产品研发、设备状态监控、生产过程优化、生产参数控制等功能。李部长此时心中充满了对企业未来的信心，决心在数据应用方面走在其他钢铁企业的前列，把握住数据先机才能赢取企业未来。

卢经理诚恳地表达了实现企业级数据挖掘系统的看法，指出对于像 T 钢铁公司这样向世界 500 强进军的企业，需要将数据挖掘的第一步走得扎扎实实：首先要对全公司的各种系统的数据库系统集成，建立企业数据仓库，为快速、有效地进行数据挖掘打好坚实的基础。

后来，T 钢铁公司与卢经理的公司全面合作，在钢铁企业建立了国内首家数据挖掘应用研究中心。

2011 年初，李部长五十岁生日也是公司企业级数据挖掘系统成功通过验收的日子。李部长深知企业级数据挖掘系统的建立是公司发展史上的里程碑，为企业更快、更好的发展注入了新的活力。

读书笔记

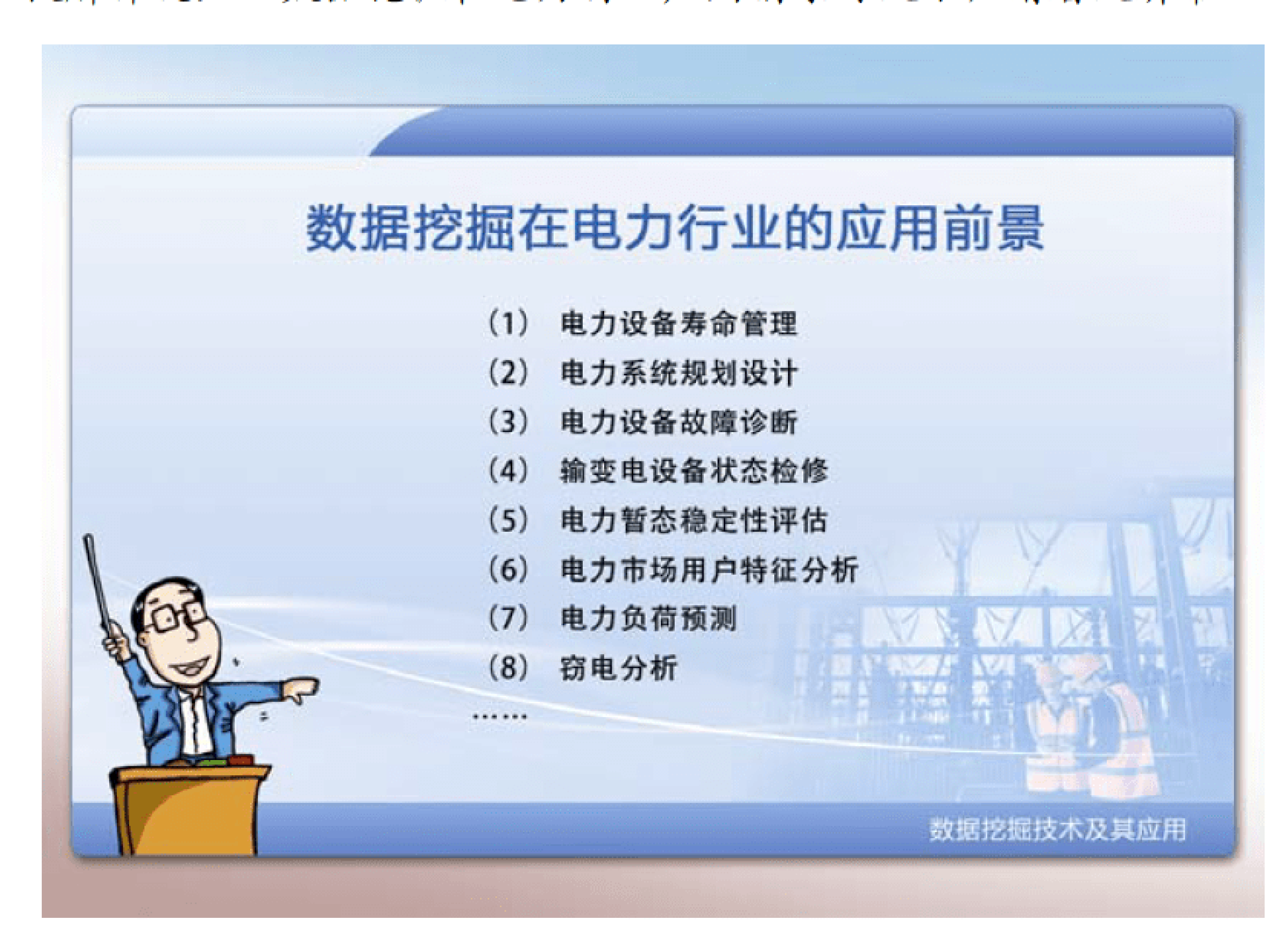
第 3 章　数据挖掘在电力行业的应用

“上一讲李部长以其亲身经历介绍了数据挖掘的流程，相信大家对数据挖掘的过程有所了解了。接下来，我们将开始一起讨论数据挖掘在电力、交通航空、冶金、银行与税务、军工等行业和故障诊断领域的应用。”徐教授边说边打开电脑。

“徐老师，俗话说近水楼台先得月，您就从数据挖掘在电力行业的应用讲起吧！”坐在教室第一排的电力公司马处长抢着说。

“数据挖掘在电力行业有什么用处呢？”有学员问。

徐教授解释说：“数据挖掘在电力行业应用前景可大了，请看大屏幕！”



3.1　应用前景

看着屏幕上的内容，马处长激动万分：“我国电力系统的信息化从 20 世纪 60 年代起步，最初主要集中在发电厂和变电站自动监测方面。20 世纪 80～90 年代各种专项业务应用系统（如电网调度自动化、电力负荷控制、计算机辅助设计、计算机仿真系统等）陆续投入使用。20 世纪末电力信息技术进一步发展到综合应用，由操作层向管理层延伸，办公自动化（OA），MIS 系统、电力市场和营销系统、能量管理系统（EMS），配电管理系统（DMS）、呼叫中心（Call Center）以及电力自动化管理系统已广泛应用。”

停了片刻后，马处长继续说：“这些系统每天都在产生大量的数据，尤其是 SCADA 系统时刻都在对现场的运行设备进行监视和控制，实现数据采集、设备控制、测量、参数调节以及各类信号报警等功能。对这些数据我们如果采用传统方法去处理，不能对其进行深层次分析而从中提取有用的信息，企业的管理水平也得不到根本提高。另一方面，我们虽明知这些数据中蕴藏着重要信息，但由于缺乏从数据库中提取有价值知识的工具，许多重要的决定通常不是根据数据库中信息丰富的数据，而是凭自身经验和直觉做出的。数据和信息之间的鸿沟我们一直无法跨越。”

马处长刚一口气讲了太多，端起茶杯猛饮了几口，接着说道：“今天，我特别高兴，我看到了徐老师总结出的电力行业连接数据与信息的这座桥梁。”

看着马处长激动地样子，徐教授说：“马处长，据我所知，我国很多省包括你们省建立了电力数据中心，一直企图利用这些数据宝库为企业的重大决策、降低成本和优化运行提供科学的、有前瞻性的决策依据。但由于缺乏数据挖掘人才而难以形成有一定实力的研究与应用团队，电力数据中心的功能至今无法得到充分发挥。”

听了徐教授的这番话，马处长脸上热辣辣的，不好意思地说：“徐老师，我们省的电力数据中心就是我一手负责建立起来的。不瞒您说，我这个兼职的中心主任

实在无法再兼下去了，人家在背后窃窃私语，说我建的是数据坟墓，还喊我马园长。”

徐教授：“哈哈，我终于明白了一件事情。在这期 EMBA 班开学前，你们省电力公司的陈总打电话问我是否要给这个班上数据挖掘课，如果是的话，他就安排马主任去学习，原来马处长就是马主任。”

“徐老师，其实我参加 EMBA 班的目的主要是为了跟您学习数据挖掘，希望得到使数据坟墓起死回生为知识宝库的法宝，扔掉园长这顶帽子！”马主任终于有了释放胸中郁闷的时机。

徐教授：“好吧，那我们就开始探讨数据挖掘在电力行业的应用，帮助马处长扔掉这顶谁都不愿意戴的帽子吧！不过，马处长，你可要与大家紧密配合。”

“当然了，大家帮我，我肯定给力了。今天晚上我请咱们全班同学去吃渭南水盆大肉。”马处长更来劲了。

徐教授将光笔指向屏幕：“马处长，我们就从数据挖掘在电力设备寿命管理中的应用开始讨论吧。请你先给大家介绍一下什么是电力设备寿命管理。”

马处长背诵如流：“**寿命管理就是在对设备进行监测和评估的基础上优化其运行和检修管理，降低设备寿命周期费用**。对于资金密集型设备，控制好寿命损耗率意义重大。寿命管理最关键最基础的要求是科学地评价材料的状态。”

徐教授点评道：“但是传统的数据分析处理方法不能有效利用现有数据而准确评价材料的状态。采用数据挖掘技术，通过神经网络、决策树、关联规则、正则化方法等手段，建立非线性预测、分类模型来研究常用材料在长期使用中的老化和损伤规律，并且把上述规律和设备的运行状态结合起来，从而提高状态评估的客观性、准确性，科学地进行电力设备的寿命管理。”

徐教授一下子讲出了这么多的数据挖掘方法，马处长一时不知所云，焦急不安地说：“这么多方法，我们啥时候能学会呀！”

徐教授安慰马处长道：“这些方法的基本思想我们在以后的课程会陆续介绍。

马处长，你是将军还是兵娃子，将军的职责是指挥作战，所以你只要明白这些方法的基本原理，成为数据挖掘的将军，内行地领导工程师们进行数据挖掘就行了。”

马处长终于松了一口气：“好，争取学完这门课程，我能够成为一位合格的数据挖掘将军。我就不纠结数据挖掘方法的实现细节了。徐老师，电力系统规划设计方面怎样应用数据挖掘技术？”

徐教授耐心地解答说：“电力系统规划设计的目的是取得有效的系统规划结果，在进行规划设计时就必须考虑由负荷模型不同引起的系统多种结构及在每种结构下可能出现的故障，由此制定出保证系统安全稳定运行的规划策略，如确定相应的临界运行参数和稳定域，确定保护和控制装置的参数。在此过程中，数据的处理量是巨大的。数据挖掘正是一个利用各种分析工具在海量数据中发现模型和数据间关系的过程, 这种模型和数据间的关系可被用来制定系统正常情况下的运行法则和发生故障时的应对策略。因此，**数据挖掘技术可被用于电力系统的规划设计**。”

马处长：“我回去后就给规划处建议，让他们立一个电力系统的规划设计的数据挖掘方法项目。”

“马处长，先别急着立项，数据挖掘在电力系统应用场合多着呢。**电力系统故障**分析也是一个很有潜力的用途方向。”徐教授说。

“徐老师，那您就赶快给我们讲讲吧！”马处长急切地想知道。

徐教授：“电力系统的故障受理系统在业务处理中积累了大量数据，可以利用数据挖掘技术将这些数据中蕴藏着的许多潜在的重要因素、事实和关联等有价值的信息提炼出来。例如，可运用数据挖掘中的关联分析法分析故障发生原因同其他因素的相关性，如故障和对其影响很大的温度、雨量、雷暴、负荷等因素之间的关系，从而使故障分析符合事物的客观规律；再运用序列模式分析方法找出几类重要的、常发生故障的、具有相同模式的部件；再按分类分析法定义出常发生故障部件的分类标准，即故障模式；最后用故障模式作为分析规则，运用聚类分析法找出在该模式下尚未发生故障的部件，作为重点预维修的参考, 实现可靠的安全管理。”

“徐老师，我看您 PPT 上还有‘**电力市场用户特征分析**’，您再给我们介绍一下数据挖掘在这方面的应用吧。”马处长极为关注这个应用方向。

徐教授：“好吧。从 1990 年开始，我国的电力系统就告别了计划经济模式，电力企业也走向了电力市场，电力用户可以选择供应方和贸易方式。因此从供电方来说，它自身是商家，而用户是消费者。在市场营销中，商家为节省营销成本获得更多的利润，应该通过收集、加工和处理消费者的大量信息确定特定消费群体和消费需求，进而推断出消费群体下一步的消费行为。然后以此为基础，对识别出来的消费群体进行特定内容的定向营销。同样在电力市场中，供电方也必须在对用户负荷的特性充分了解的基础上，对用户的行为分门别类，从而在保证系统安全稳定运行的前提下，制定出有竞争力的供电策略。考虑到电力系统自身特点，供电方还应制定有效的负荷管理策略，调整负荷曲线的形状，降低对峰荷的要求，节约能源。上述工作都可以采用数据挖掘技术进行。”

马处长越听越觉得糊涂，直截了当地大发感慨：“徐老师，数据挖掘在电力行业真是太有用武之地了，可就是用到的数据挖掘技术太多，我真是丈二和尚摸不着头脑！其他同学也应该与我差不多。”

徐教授温和地说：“大家不用着急，我为你们准备了大量行业领域数据挖掘的应用实例，以后课程且听我慢慢分解，相信你们逐步会明白数据挖掘的奥秘！”

李部长知道，徐教授肩负 973 首席科学家的重任，科研任务那么繁重，还对数据挖掘这么“平凡”的课程付出如此巨大的心血，情不自禁道：“徐老师，您为我们上课花了太多的功夫，我们大家真是过意不去。”

李部长这么一说，徐教授也激动了：“我们国家在数据挖掘研究方面可以说基本与国际同步，甚至有些研究领域处于国际领先地位，但是我们在数据挖掘的应用方面，普及率太令人不安了。在这门课的开场白中，我不是希望大家一起为建设创新型国家做贡献吗！我觉得，我这样的付出非常值得，因为这就是我的具体行动！”

徐教授话音刚落，全体学员都站了起来，一阵长时间的掌声，淹没了下课的铃声。

3.2 电力设备状态检修

上课铃响了，徐教授径直走上讲台说：“这一节，我们一起讨论数据挖掘技术在电力行业的应用，你们会越来越发现数据挖掘的无穷魅力。首先我们探讨数据挖掘在电力设备状态检修中的应用。马处长，你是电力公司管设备的，给大家介绍一下什么是状态检修吧。”

马处长依旧坐在第一排，站起来说：“唉，不怕大家笑话，因为前段时间一台330KV 变压器出现了故障，导致大面积停电，我已被降为副处长了。大家还喊我马处长，我也不好意思解释。”

李部长为马处长打抱不平：“马处长就是马处长，你就是被降为科长我们大家还叫你马处长。在你们电力公司谁人不知，你以公司为家，任劳任怨，敢于担当风险，不计个人得失，干得越多出差错的可能性就越大。”

S 钢铁公司的赵总也开了口：“马处长，别气馁！你们省电力公司的陈总叫你来 EMBA 班学习的目的是司马昭之心路人皆知，他希望你从哪儿跌倒从哪儿爬起来。从我们这个班毕业，你的管理能力当然会提升到一个新的高度。重要的是，跟徐教授学好了数据挖掘，并在你们电力行业付诸应用。特别是电力设备状态检修方面，只要勇敢地实践，在国内做到首屈一指，并努力赶上甚至超过国际先进水平。这样，马处长的宝座，不，是副总的位子自然非你莫属了。”

马处长摆了摆手：“别拿我开心了，我还是开始给大家介绍什么是状态检修吧。”

“马处长，好像在状态检修应用之前很长时期电力部门实行的是定期检修吧？”李部长问。

马处长抱怨道："定期检修，累死人了。我才毕业参加工作那几年，实行的是计划检修，我师傅带我一年四季都在忙于设备检修。长期以来，我国对电力设备的检修策略主要采用以时间为标准的定期维修。虽然定期维修一般可在维修时发现设备存在的缺陷，对保证设备安全运行发挥了重大作用。但是，定期维修存在'维修过剩'和'维修不足'的缺陷。不仅造成部分设备盲目检修，导致人力和物力的大量浪费，而且增加了产生新隐患的几率，降低了供电可靠性。"

贾总经理说："马处长，大家都知道定期检修已经成为历史了，在当前电力企业走向市场的形势下，用状态检修的模式代替传统定期检修制度是电力企业自身发展的必然趋势。那么状态检修到底是怎么一回事?"

绕了这么一大圈，马处长终于切入主题：**"状态检修是以设备的当前实际工作状况为依据，通过先进的状态监测手段、可靠的评价手段和寿命的预测手段来判断设备的状态,并识别故障的早期征兆，从而根据分析诊断结果在设备性能下降到一定程度或故障将要发生之前进行维修。"**

听了马处长对状态检修的介绍，李部长抓住了问题的关键："这么说状态评估是电力设备状态检修的基础，状态评估采用什么方法呢？"

徐教授概括道："随着状态检修理论的研究与应用，设备状态综合评估技术得到了国内外研究机构和电力企业的深切关注，纷纷展开对电力设备状态评估方法的研究，但仍处于探索阶段。近十几年来，数据挖掘的新方法不断涌现，为设备状态评估提供了新的思路。采用数据挖掘技术，对设备的监测、试验数据进行分析，揭示电力设备性能状态渐变和寿命损耗的规律，及时、准确地发现潜在故障的早期征兆，快速对故障部位严重程度及发展趋势做出判断，确定科学的检修计划。"

李部长听完，似乎受到了一些启发，问道："在电力系统，一般对哪些设备进行状态检修？"

马处长不假思索地答道："这可多了，发电厂设备、变电站设备、输电线路和配电设备，只要能够反馈有效工作数据的重要设备都可以施行状态检修。"

李部长即刻兴奋起来："我明白了，电力设备的状态检修方法可以推广到钢铁、化工、铁路、航空航天、制药、电子等制造型企业，数据挖掘技术可大有用武之地了。"

徐教授："这几年，各行各业都开始关注数据挖掘技术的应用，到我这儿咨询的人越来越多。"

马处长："徐老师，您能不能以一种典型设备为例，详细讲解一下应用数据挖掘技术进行设备状态检修的具体方法？"

"好吧，变压器作为输配电网的主要枢纽设备，其安全可靠性尤为重要。我们就一起探讨变压器状态检修的数据挖掘方法吧。"徐教授说。

"变压器状态检修的原理是什么呢？"李部长问。

徐教授不紧不慢地说："大家都知道变压器的长期发热使得矿物绝缘油和固体有机绝缘材料逐渐老化、变质，在这个过程中伴随产生各种气体，当然这个是一个缓慢的过程。"

"也就是一个由量变到质变的过程。"台下有人附和。

徐教授说："不错！当变压器内部发生故障时，由于电、热故障的结果使某些C-C键和C-H键断裂，伴随生成少量活泼的氢分子和不稳定的碳氢化合物的自由基，这些氢原子或者自由基通过复杂的化学反应迅速重新化合，形成一些气体，如氢气（H_2）、甲烷（CH_4）、乙烷（C_2H_6）、乙烯（C_2H_4）、乙炔（C_2H_2）、一氧化碳（CO）和二氧化碳（CO_2）等。随着故障的日益严重，相应的气体浓度不断增加。"

"哦，这就是反映变压器状态由量变到质变的过程喽！"台下有人嘀咕道。

徐教授接着说："在故障初期，所形成的气体溶解于油中。当故障能量较大时，这些气体的产量会快速增加，故障点温度较低时，CH_4 比例较大；温度升高时，C_2H_4、H_2 成分急剧增加，比例增大。当严重过热时，则会生成 C_2H_2 气体。当发生固体绝缘过热性故障时，除产生上面的低分子烃类气体外，还会产生较多的 CO 和 CO_2，

且随着温度的升高，CO/CO_2 的比值逐步增大。因此，可以通过定期测量变压器油中的各种气体含量，应用相关的气体分析技术，判断变压器故障的性质和程度，为状态检修安排提供依据。”

马处长插话说：“在变压器状态检修技术开展的初期，主要是通过分析油中溶解气体的含量及相互关系对变压器进行状态诊断的方法，那是比较流行的方法是三角图法、三比值法等。”



徐教授补充说：“这些方法大多仍局限于阈值诊断的范畴，一般只给出一个判定边界的描述。这样难以确切反映故障与表现特征之间的客观规律，并且很难在溶解气体含量较小的情况下对变压器状态进行分析。也就是说，只有当某些特征气体含量超过“临界值”时，判断结果才被认为是有意义的。传统方法的这些缺点无疑对变压器潜伏性故障的发现和分析非常不利。而数据挖掘技术应用于变压器状态检修可显示出很多优点。”

“徐老师，您一定把这部分讲解详细些，变压器状态检修可是当前国家电网很重视的一个研究和应用热点！”马处长说。

徐教授：“好，我在讲数据挖掘在电力行业的应用时，会顾及到其他行业的学员，从基础的东西开始讲解。希望能够抛砖引玉，使大家借鉴到自己的行业中，解决本行业的具体问题。”

马处长对这节课非常感兴趣，催促道：“徐老师，咱们赶紧开始讲数据挖掘技术在变压器状态检修中的应用吧？”

徐教授一语道破了变压器状态评估的基本思路：“我们应用支持向量机回归方法，对反映变压器状态的各种因素，建立变压器状态回归模型，从而可以根据这些因素的变化来快速评估变压器运行状态。”

“为什么采用支撑向量机技术呢？它有什么优势呢？”台下有人问。

徐教授回答道：“支撑向量机英文为 Support Vector Machines，简称 SVM，是数据挖掘中的一项非常有效地解决分类和回归问题的方法，最初是 20 世纪 90 年代 Vapnik 等人根据统计学习理论中结构风险最小化原则提出的。该方法在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势，所建立的模型具有简洁的数学形式和直观的几何解释。且求解算法需要人为设定的参数少，便于应用，得到的模型具有良好的泛化能力，为小样本机器学习提供了一种新方法。”

马处长：“这么吸引人的方法，徐老师，可否给我们具体讲解一下支撑向量机的建模过程？”

徐教授：“我给研究生上课时，支撑向量机的建模过程会讲解得非常详细，要给你们这样讲，肯定大多数人就会眼睛一闭一睁一节课就没了。如果确实有人对此感兴趣，我们课后可抽时间讲解。”

“好吧，徐老师，今晚我们对此感兴趣的几个人到您办公室，您给我们开个小灶吧。”马处长恳求道。

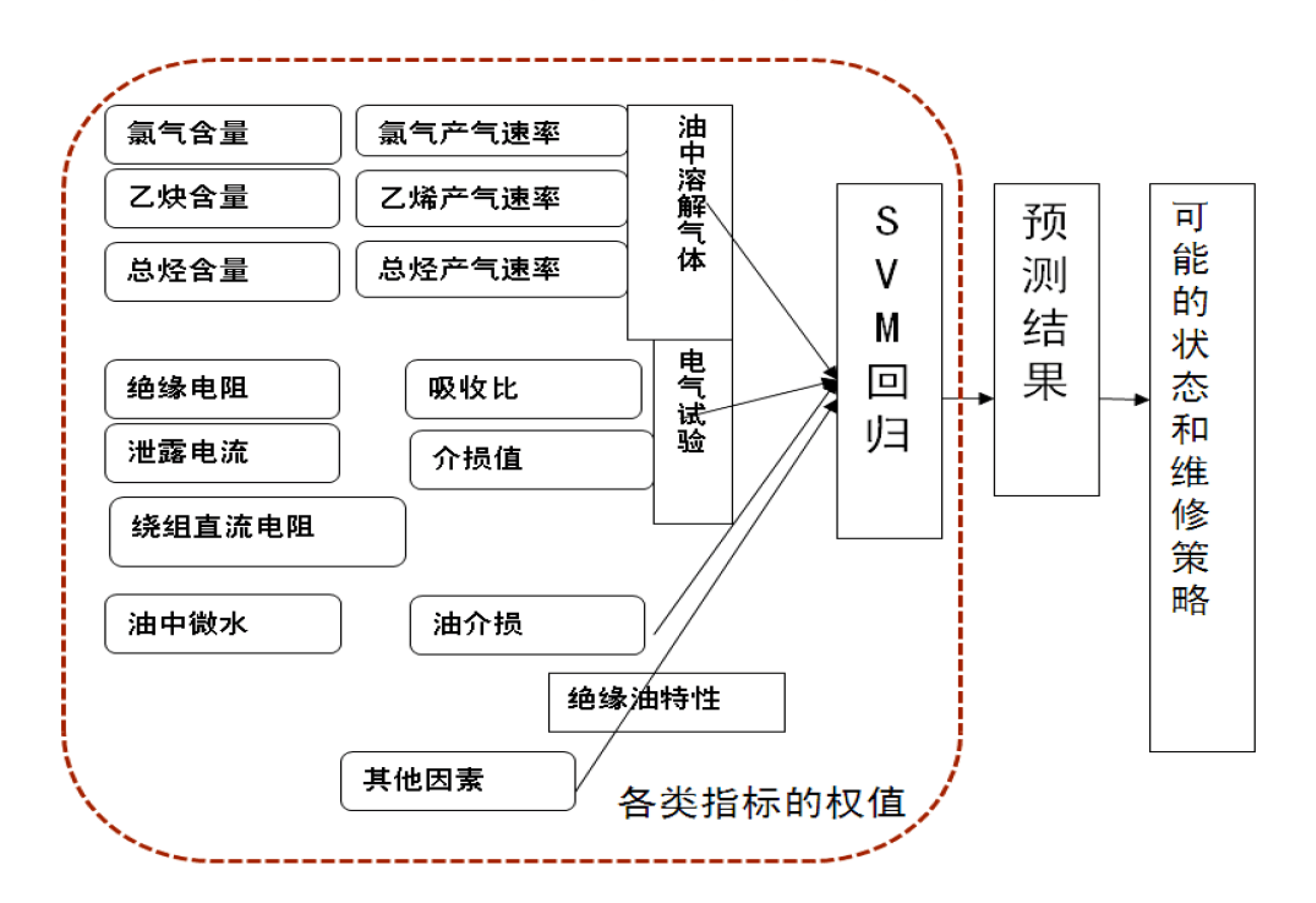
“没问题。”徐教授干脆地应道。

“徐老师，听我们公司信息中心的小张说，现在的数据挖掘平台如 Clementine、Weka，还有国产的 Merit DM 等都集成了支撑向量机方法，只要按要求将数据代入就可得到预测模型。”

“是的，大家只要学会了这些软件的使用，就可以应用支撑向量机方法解决实际问题了。”徐教授答道。

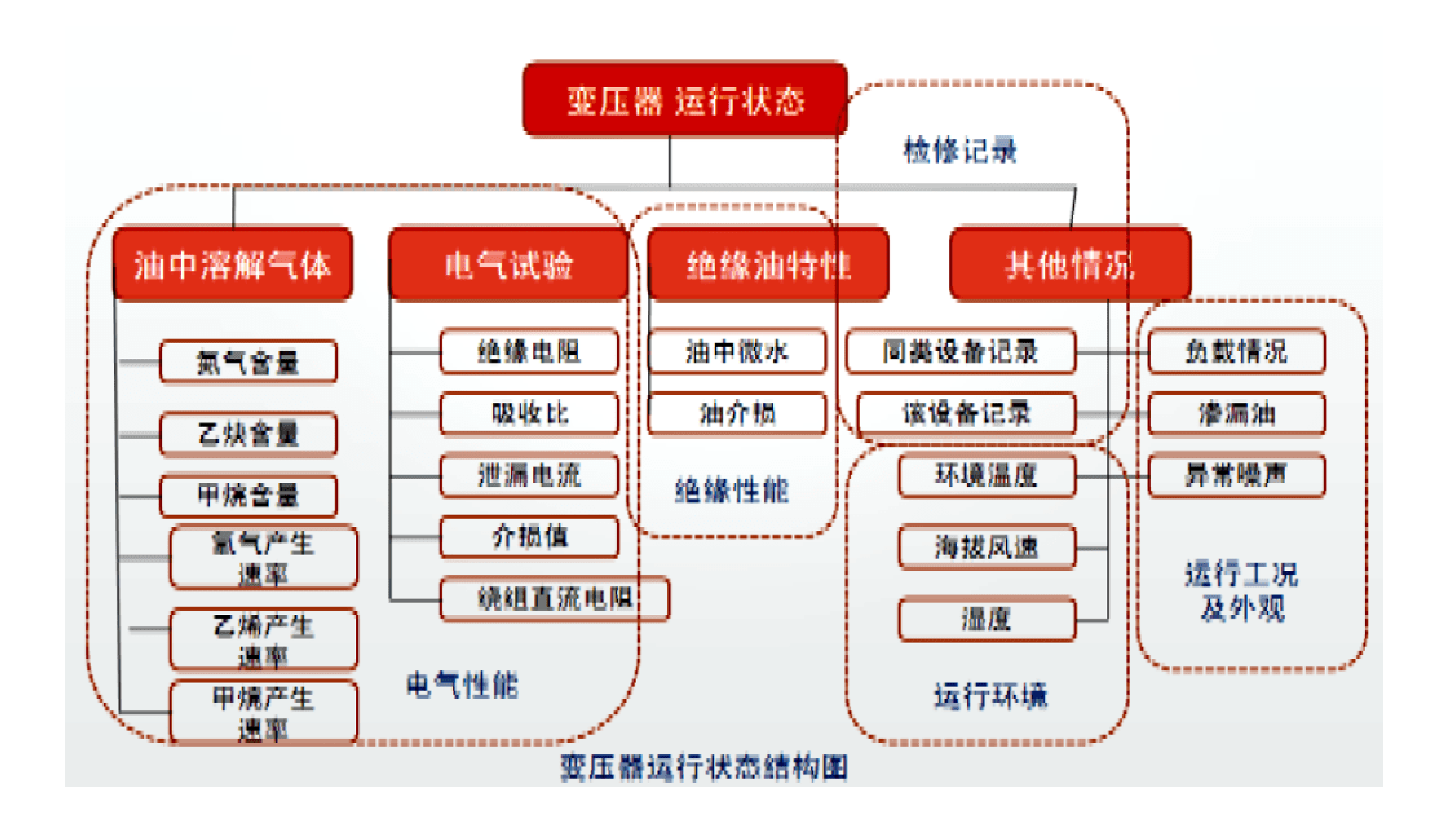
“太好了，徐老师，那您就给大家讲讲如何运用支撑向量机来建立变压器状态评估模型吧！”马处长又一次发出了请求。

徐教授：“我们采用支撑向量机回归技术，将对变压器的各个评价指标作为输入，将实际评估结果作为输出，通过对训练集进行学习，在测试集上对得到的模型进行测试，最后得到预测能力强的变压器状态评估回归模型。”



看着屏幕的内容，马处长说：“我看明白了，变压器的评价指标都与变压器状态有关，或者说能够直接或间接反映变压器状态，例如油中各种气体的含量、电气性能、运行环境等。”

“马处长说得比较笼统，但基本都概括了。具体的评价指标大家可以看下一张PPT，我把各类指标细化了，有什么不明白的可以问马处长。”徐教授说道。



“哦，我想问马处长一个问题，这些评价指标的数据通过哪些途径获取？”

马处长如数家珍地说：“这些数据主要来自 SCADA（数据采集与监视控制系统）、EP-MIS（电力流程化管理信息系统）、GSRMS（电网空间资源管理系统）、PIMS（生产实时信息管理系统），还有日常的实验数据等。”

“这么说电力公司还未建立数据仓库？”李部长问道。

马部长转向李部长解释说：“没有。数据挖掘在我国的电力行业只有零星的应用，虽然有些电力公司建立了数据中心，但由于对数据挖掘认识不够，数据中心并

不是真正意义上为数据挖掘服务的数据仓库。”

李部长感慨万分：“那么在变压器状态评估的数据准备阶段就要花很大的力气了，这样会大大影响数据挖掘的效率和质量。”

“是的。我回去建议我们公司率先建立数据仓库，李部长，你们 T 钢铁公司也尽快建吧，我们一起交流、协作，共同进步吧。”马处长道。

李部长：“好吧，我们达成口头‘君子协定’，回到公司我们就开始起草建立数据仓库的建议书。”

马处长不加思索地回应道：“一言为定！”

S 钢铁公司的赵总有点不耐烦了：“你们俩下课后好好商量，不要浪费课堂时间了。我这里有数据仓库专业公司程总的电话，你们去向他咨询吧。”

徐教授对他俩的想法表示肯定，并鼓励道：“对你们这些实力雄厚的大公司，建立数据仓库确实是正确的选择，但愿你们能够成功，为数据挖掘在我国走向应用、赶上甚至超过国际先进水平探索道路。”

徐教授将 PPT 翻了一页，继续道：“建立变压器状态评估的回归模型，在选定了输入指标后，还要有一定的对评价结果进行度量的策略，请看变压器状态检修评估结果表：”

评分	0～20	20～40	40～60	60～85	85～100
状态	严重	异常	注意	良好	优秀
维修策略	立即维修	尽快维修	优先安排	计划	延期
描述	单项重要状态严重超过标准限值，应立即维修；	单项重要状态变化较大，应监视运行；	单项或者多项状态变化阶级标准，仍可以运行；	各状态处于规程的警示外，比出厂状态有一定差距；	各变量处于稳定，性能接近出产值；

徐教授接着讲：“设备变压器的健康状态评分，分值为 0～100 分，0 分表示设备

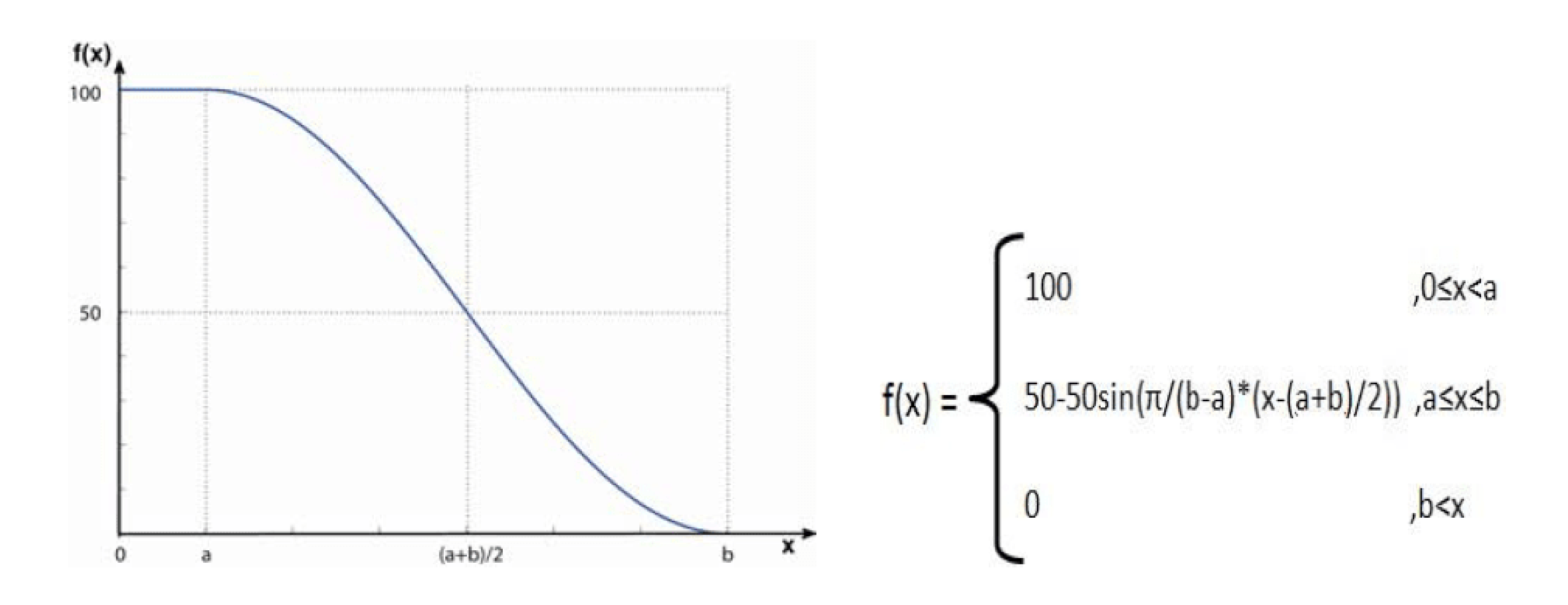
需要马上检修；100 分则表示变压器的各项指标都很正常，而且没有经历不良工况，又没有家族质量缺陷史，即设备完全处于正常状态，无需维护。有了这样的一个较为详细的评价体系，我们就可以开始数据挖掘下一阶段的工作了。”

“下一步就是收集数据，并进行数据预处理了吧？”马处长问。

徐教授回答道：“不错！马处长已经对数据挖掘流程非常熟悉了！数据准备阶段要对数据质量进行评价，然后对噪音数据、缺失数据等进行处理，为数据建模打好基础。”

“徐老师，上面您讲过，我们要建立变压器评估模型的输出为对变压器综合状态的百分制评分，我想是不是对对变压器的各个状态指标也要以百分制评分？”马处长道出了自己的思路。

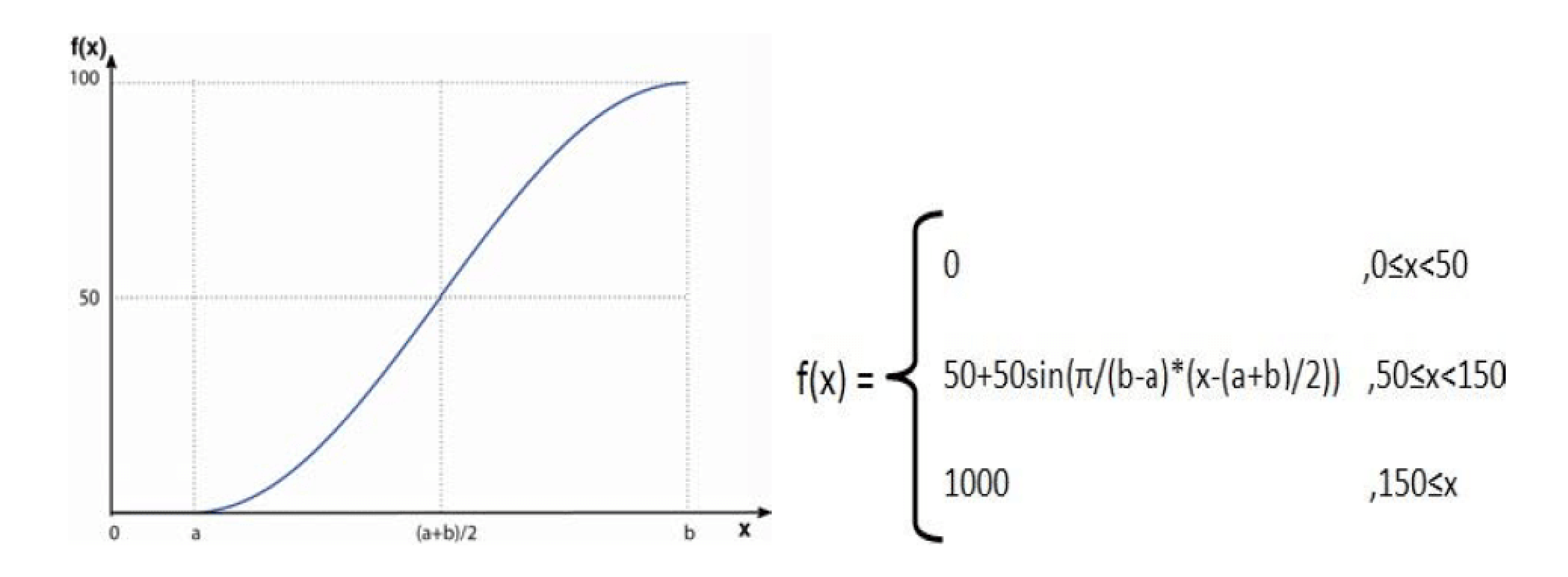
徐教授肯定地说：“马处长不愧为设备管理出身，一语道破了研究者多年探索才得出的方法。我们采用半岭模型来对变压器的各个状态指标进行单个评估。请大家看屏幕，其中，a 和 b 是变量的阈值，x 是评分参数的实际测量值，f（x）为评分的结果值。对于数值越大越好的指标，采用升半岭模型，反之采用降半岭模型。”



徐教授指着投影屏幕继续说：“降半岭的评分公式如下，它为取值在 0～100 之间的单调下降函数。”

“同样升半岭的评分公式如下，它为取值在 0～100 之间的单调上升的函数。”


徐教授接着说。



“徐老师，评分公式很简单，相信大家都可以看明白。具体在变压器各个状态指标评估中如何应用呢？”马处长希望知道的更详细。

徐教授说：“好，那我就举些具体的例子吧。首先我们看看对变压器特征气体指标的评分。在变压器状态检修中围绕如何确定特征气体的含量与变压器内部故障的定量关系，国内外的 DGA 技术工作者都做了大量的研究。在统计结果的基础上，DL/T772-2000《变压器油中溶解气体分析和判断导则》、《油浸式变压器（电抗器）状态评价导则》和其他参考资料给出了变压器油中气体含量的推荐阈值，并依据指标的优劣采用不同的半岭模型评分规则。大家请看屏幕，显示的是变压器状态气体评分模型与阈值表。我们可以根据该表完成对特征气体指标的评分。

特征气体	A 出厂值（ul/L）	B 注意值（ul/L）	评分模型
H_2	50	150	降半岭模型
C_2H_2	0	5	降半岭模型
总烃	20	150	降半岭模型

“同时我们制定出变压器绝对产气速率注意值与阈值表，就如屏幕显示。”徐教授指着屏幕说。

特征气体	A 出厂值（ul/L）	B 注意值（ul/L）	评分模型
H_2	5	10	降半岭模型
C_2H_2	0.08	0.2	降半岭模型
总烃	8	12	降半岭模型

“具体对每个指标的评分方法是怎么样的？”马处长问。

“好，以 H_2 为例，来说明一下 H_2 含量评分函数和产气速率评分函数。”徐教授说。

$$f(x)=\begin{cases}100 & ,0\le x<50\\ 50-50\sin(\pi/100(x-100)) & ,50\le x<150\\ 0 & ,150\le x\end{cases}$$

$$f(x)=\begin{cases}100 & ,0\le x<5\\ 50-50\sin(\pi/5(\pi-7.5)) & ,5\le x<10\\ 0 & ,10\le x\end{cases}$$

“首先我们分析指标特征后，觉得采用升半岭评分模型还是降半岭评分模型，然后把阈值（也就是我们所说的注意值）带入模型就可以了。”徐教授说。

“那么对于变压器绝缘性能指标是不是也可以这么做？”马处长又抛出个问题。

徐教授回答：“是的，以绝缘电阻为例，绝缘电阻指的是在绝缘结构的两个电极之间施加直流电压与流经该对电极的泄漏电流值之比。变压器的绝缘电阻越高表示绝缘性能越好，根据《规程》，220KV 及以下的变压器其绝缘电阻在 20℃时不小于 800MΩ，若大于 1600MΩ 表示绝缘电阻状态良好，所以评分模型采用升半岭模型，其评分函数如屏幕显示。”

$$f(x) = \begin{cases} 0 & ,0 \le x < 800 \\ 50+50\sin(\pi/800*(x-1200)) & ,800 \le x < 1600 \\ 1000 & ,1600 < x \end{cases}$$

“那么对于检修记录如何量化呢？”马处长又问。

“马处长你真是成了问题篓子喽。”台下有人调侃马处长。

徐教授说：“问题多说明善于思考！马处长的这个问题很重要。对于检修记录的量化，变压器的检修历史、运行环境、外观检查以及部分运行指标属于定性指标，度量困难，需依靠专家经验进行定性描述，故需要专家打分，范围是[0,100]，若指标反映变压器状态越好，分值越接近 100。”

评分	0～20	20～40	40～60	60～85	85～100
检修记录	难度大，次数多，有明显缺陷	难度偏大，次数偏多，有一般缺陷	难度一般，次数一般，有一般缺陷	难度一般，次数较少，留轻微缺陷	难度偏小，次数偏少，未留缺陷
家族史	难度大，次数频繁	难度偏大，次数偏多	难度一般，次数一般	难度一般，次数较少	难度偏小，次数偏少

徐教授接着说：“例如，对运行中变压器遭受的过电压的量化要根据遭受过电压的大小和次数来获得指标评估量化值，遭受过电压最大电压越高，次数越多，则值越小。”

“那么对于外界环境因素如何量化呢？”马处长问。

“变压器的运行环境一般包括：周围空气的温度、湿度、污秽等级别，考虑的环境因素记录等问题采用温度和湿度进行量化。”徐教授解释道。

环境指标	0～20	20～40	40～60	60～85	85～100
温度（摄氏度）	年平均温度>20 度，极限温度经常>40 度或<-25 度	年平均温度>20 度，极限温度有时>40 度或<-25 度	年平均温度接近 20 度，极限温度有时>40 度或<-25 度	年平均温度接近 20 度，极限温度偶尔>40 度或<-25 度	年平均温度地域 20 度，极限温度没有>40 度或<-25 度
湿度	>90%	80-90%	60-80%	40-60%	<40%

“那么对于变压器外观的量化呢？”马处长又问道。

徐教授说：“我们对于变压器外观指标的量化主要是通过漏油的严重程度来划分的。如屏幕所示的标准来进行外观的量化。”

外观指标	0～20	20～40	40～60	60～85	85～100
渗漏油	漏油	多处明显渗漏油	一处明显渗漏油	多处轻微渗漏油	一处轻微渗漏油
异常噪音	声音增大并有明显杂音	声音变高，夹杂有尖锐声	声音均匀，比正常时沉重	声音均匀，比正常时稍大	声音平稳，均匀

李部长通过数据挖掘成功地解决过硅钢质量控制问题，他知道下一步该干什么了，便说：“输入输出数据都有了具体的量化方法，便可以收集数据，训练出变压器的评估模型。”

“是的。某省电力公司三年一次的安全性评价开始了，他们从各地区供电公司遴选了 15 位专家，巡回对全省 19 个地市的 216 台型号为 SFPSZ8-120000/220 的变压器进行了安全性评估，收集到了详尽的评估数据。”徐教授说。

李部长曾经领导过几个数据挖掘项目，对 SVR 回归方法步骤比较熟悉，他说：“我想，对数据集进行标准化处理后，在应用 10 倍交叉验证法对 SVR 模型进行训练，就可得出变压器状态评估模型。”

徐教授：“李部长说得很对，就是这样的步骤。训练模型的过程，也就是支撑向量回归机对专家评估经验学习的过程。”

马处长豁然开朗：“徐老师，您看我的理解对不对。支撑向量回归方法假定了一个学习机器（即带有一些参数的函数表达式），机器学习就是利用数据集反复训

练找到学习机器中最优的参数，从而使学习机器变成一个具体的回归函数表达式，这个回归函数对未来数据具有较好的预测能力。具体的说，对变压器评估的数据集，应用支撑向量回归方法训练得到了一个囊括众多评估专家经验的回归函数，以后对型号为 SFPSZ8-120000/220 的变压器，将收集到的变压器状态数据代入这个回归函数，就可以知道这台变压器可以得多少分，据此确定对其状态检修的措施。”

听着马处长的表述，徐教授不断点头。他刚一说完，徐教授便给予肯定地回答：“马处长理解地完全正确！”

“徐老师，通过支撑向量回归方法，得到了变压器评估模型以后，以后在实际工作中怎么应用呢？”马处长脑子里又闪现出了一个问题。

徐教授不加思索，脱口而出：“是这样，Merit DM 数据挖掘平台中，可以将学习得到的模型固化，各地市供电公司安监科或变电站直接使用包含着众多专家智慧的回归模型对变压器进行评估。这样，就好像专家们成了随时都可以请到的顾问。”

听到这里，马处长好像取得了真经，高兴得站了起来：“不错，这样，使得状态检修工作可以常态化，真正做到防患于未然。”

“马处长，你只顾高兴，其他学员对支撑向量机应用于状态检修不一定理解的与你一样透彻。不过，我们以变压器状态检修为例，起到一个抛砖引玉的作用。其实数据挖掘技术在很多行业的关键设备上的应用都能够如法泡制。”徐教授总结道。

一时，教室里激烈讨论开了，大家纷纷谈论自己行业哪些方面可以应用数据挖掘技术进行状态检修。

这时下课铃响了，徐教授边收拾笔记本电脑边说：“好，大家好好聊吧，相信数据挖掘在你们的工作中会有用武之地的。”

3.3 电力系统暂态稳定性评估

徐教授打开幻灯片，说："上节课我们介绍了电力设备状态检修，今天我们来讲下电力系统暂态稳定性评估。"

"电力系统暂态稳定性？"大部分学员根本没听过这个名词，有人疑惑不解地问道。

"这个问题，就由电力公司的马处长给大家介绍一下吧，他是专家！"徐教授说，然后端着杯子坐到了教师第一排座位上。

"马专家，该你上台了！"坐在马处长旁边的李部长，拍了拍马处长，鼓励他说。

马处长清了清嗓子说："电力系统在大扰动后，如发生各种短路故障、切除大容量发电机、输电设备或某些负荷的突然变化等情况，如果电力系统能够保持同步运行，并具有可以接受的电压和频率水平，则称此电力系统在这一大扰动下是暂态稳定的。在电力系统规划、设计、运行和控制时都要进行大量的暂态稳定分析。通过暂态稳定分析还可以研究各种稳定措施的效果以及稳定控制的性能，因此对电网的安全运行有着非常重要的意义。"

"那么，进行电力系统暂态稳定性评估目的为了什么呢？"李部长问道。

马处长说："电力系统是一个复杂的动力系统，其复杂性表现在一方面必须保证必要的电能质量及数量；另一方面系统又处于不断的扰动之中，并且扰动发生的时间、地点、类型及其严重性都是随机的。在扰动发生后的系统动态过程中，一旦发生稳定性破坏，系统可能产生严重的后果，造成极大的经济损失及重大的社会影响。对暂态稳定性进行评估，可以采取相应措施，避免电力系统故障，减少经济损失。"

"原来是这样，那扰动和你经过树林，鸟群的'躁动'一样吗？"幽默的李部长把大家全逗乐了。

马处长，笑着说："哈哈，肯定不一样喽！人给鸟一个扰动，鸟飞了，可以飞远，不再回到刚才的嬉戏玩闹的状态。而电力系统受到干扰后，必须立即采取有效措施，尽快达到新的稳定状态。"

"那扰动后的电力系统会出现哪些情况？"台下有人问。

马处长回应说："扰动后的暂态过程可能有两种不同的结果，一种是发电机转子间相对角度随时间的变化呈摇摆状态且振荡幅值逐渐衰减，各机组之间的相对转速最终衰减为零，使系统回到稳定前的稳态运行点，或者过渡到一个新的稳态运行点。在此运行状况下，所有发电机仍然保持同步运行，这样的电力系统是暂态稳定的。"

"那另一种情况呢？"李部长说。

"另一种结果是暂态过程中某些发电机转子之间的相对角度随时间不断增大，它们之间始终存在着相对转速，使这些发电机之间失去同步。发电机间失去同步后，将在系统中产生功率和电压的强烈振荡，会使一些发电机和负荷被迫切除，在严重的情况下，甚至导致系统的解列和瓦解。这种情况电力系统是暂态不稳定的，或称电力系统失去暂态稳定。"马处长接着说。

"马处长还是有两把刷子的！"台下有人说。

马处长接着说："根据在扰动后的不同时间里系统各部分的反应不同，在分析暂态稳定时往往分为三个阶段，分别是起始阶段、中间阶段和后期阶段。"

"这三个阶段有什么具体含义呢？"李部长问。

马处长解释说："起始阶段即故障后约一秒钟内的时间段。在这期间系统中的保护和自动装置有一系列的动作，例如切除故障线路和重合闸，切除发电机等。在这个时间段中发电机的调节系统还来不及起到明显的作用。"

"那中间阶段呢？"坐在前排的刘经理问道。

马处长回答说："中间阶段是在起始阶段后，大约持续 5 秒钟的时间段。在此期间发电机的调节系统将发挥作用。"

“哦，中间阶段比起始阶段时间稍长了点！”细心的李部长嘀咕到。

“后期阶段是指在故障后几分钟内这段时间。这时热力设备（如锅炉等）将影响到电力系统的暂态过程，另外，系统中还将发生由于频率的下降自动切除部分负荷等操作。”马处长说。

“后期阶段就是根据扰动状况采取措施了。”李部长说。

“不错！基本的业务知识我就介绍到这了，剩下的具体用数据挖掘技术来做电力系统暂态稳定性评估还要请徐教授来给大家讲吧！”马处长说着走下讲台。

徐教授回到讲台上说：“马处长讲得很详细、很具体！下面咱们来学习下基于数据挖掘技术的电力系统暂态稳定性评估。”

马处长赶紧拿出笔记本开始记录。

徐教授：“在电力系统运行方式变化时，经验丰富的现场运行人员常可粗略地预测出某些状态量，如母线电压、线路潮流等。这是因为运行人员通过长时间的运行，掌握了代表电力系统安全运行水平的关键部位的状态量和其他一些量的关联关系，他们可根据电力系统中控制量和扰动量的变化趋势，预测出这些关键部位在运行方式变化时的状态量，这在很大程度上是一种经验的积累。若要将这种积累以数学的形式表示出来，数据挖掘确实是最好的一种选择。”

听着听着，马处长的眼珠直打转，一个新的问题蹦了出来：“利用数据挖掘处理暂态稳定问题，有什么过人之处？”

徐教授：“数据挖掘以其自身的黑箱子特性，适合用于处理电力系统暂态稳定评估这样复杂的非线性问题。它的优势在于不受电力系统复杂的数学模型的限制，可以形成直观的规则以指导人们在电力系统暂态稳定评估中进行决策控制，为防治重大事故的发生提供更好的理论依据。”

马处长步步紧逼："徐老师，我们具体应用什么数据挖掘方法进行暂态稳定评估呢？"

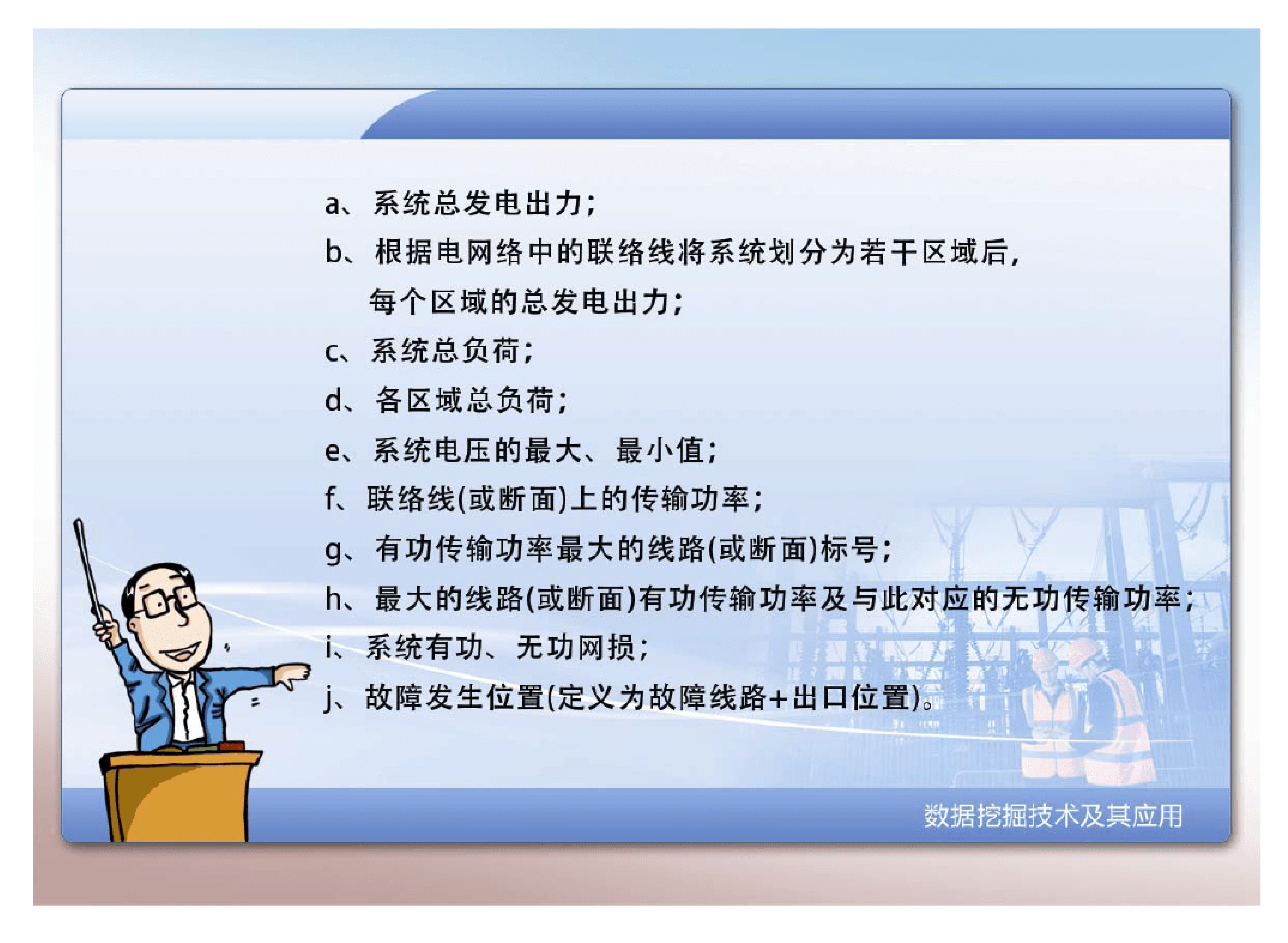
徐教授："提到数据挖掘的使用技术，就不能不提到关联规则。关联规则的发展是数据挖掘中最成功和最重要的方法之一，也是当今数据挖掘中一个非常活跃的研究领域。由于关联规则挖掘可以发现用传统方法无法发现的项与项或属性与属性间的关系规律，因此具有重要的研究与应用价值。电力系统暂态稳定评估中的关联规则主要体现在从海量数据中发现属性与属性间的频繁模式、相关性或因果关系，以便从宏观上把握电力系统所有组成元素间的关联特性。例如，在考虑所有可运行方式下，数据属性参数的变化与系统安全稳定程度之间的关联规则。"

这时，李部长也展开了联想，并表述自己的想法："徐老师，我想，进行暂态稳定评估，首先需要确定问题的变量集，并且还要考虑到在线评估对计算速度的要求，所以选择的变量不宜过多。"

徐教授又开始调动马处长的头脑中的电力知识系统："马处长，你从事电力暂态稳定性研究多年，还是给大家介绍一下哪些数据或者统计值可以作为暂态稳定性评估的特征变量？"

马处长挠了挠头，然后果断地回答："好吧，我简单说一下。不过我得从网上下载一下以前的一个 PPT。"

马处长用徐教授的笔记本登录自己的 FTP 服务器，打开了 PPT，翻出如下页面：



马处长指着 PPT 屏幕说：“这些变量和统计值，就是基本的暂态稳定性评估的特征变量，这里的出力和负荷均包括有功、无功两个部分。”

徐教授将光笔指向 PPT 屏幕，补充道：“其实还有其他一些变量，我们这里不再细究。”

突然，李部长惊叫起来：“徐老师，我记得关联规则只能处理离散数据，而我们这里选取的变量大都是连续型的。”

徐教授将目光移向李部长：“你说得很对，对暂态稳定性评估数据进行关联规则挖掘的主要难点之一就是将连续属性数值离散化。”

马处长又追问具体细节：“那到底怎样对连续数据离散化的？”

徐教授："具体地说，比如采取聚类算法找出候选离散断点，再结合信息熵理论确定最终离散断点，将连续数据离散化到各个离散区间中，然后把离散化区间映射为连续的数字标识。"

徐教授看到好多学员疑惑不解地样子，继续说："你们只要知道，连续型数据可以通过各种方法离散化，使其适应处理离散型数据的算法就行了，数据挖掘平台软件一般集成有这些数据预处理方法。"

李部长又发现了一个问题，这回他显得非常平静："徐老师，您给我们讲过，关联规则挖掘通常都是针对单维数据库，其经典的 Apriori 算法是一种在单维数据库中找频繁项集的单维关联规则算法，而我们遇到的暂态稳定数据集是多维的，这又如何处理？"

徐教授未曾预料到有学员会提出如此深刻的问题，"我们通过对 Apriori 算法（包括它的一些变形算法）进行了深入研究，然后将其改进使之适用于多维关联规则挖掘。这样便可找出暂态稳定特征属性之间以及特征属性与暂态稳定属性类别之间的关联关系。最后对挖掘出的规则进行分析研究，从而对电力系统暂态稳定评估及预测提供决策规则。"

马处长："徐老师，您具体给我们展示一些改进的 Apriori 算法发现的一些关联规则吧，以便我们对挖掘出来的规则有直观的理解。"

徐教授将 PPT 翻了几页，说道："屏幕上的规则是我们对某电力系统运用改进的关联规则方法进行暂态稳定评估所得出的几个典型的规则。"

规则1 (支持度为42. 593%，置信度为100.000%)：
当联络线路2上从发电机节点GEN2-230到负荷节点STNA-230的线路有功传输功率小于0.932 495时，系统稳定；

规则2 (支持度为62. 963%，置信度为99.020%)：
当切除故障线路后得到的IZ≥-0.017 55时，系统易稳定；

规则3 (支持度为4. 321%，置信度为85.714%)：
当故障发生在联络线路2靠近发电机侧，且联络线路2上从发电机节点GEN2-230到负荷节点STNA-230的线路有功传输功率大于0.932 495时，系统易失稳。

数据挖掘技术及其应用

徐教授接着解释说：“从挖掘的结果可以看出，当联络线路靠近发电机侧发生故障时，系统失稳概率较高；当在非联络线路且远离发电机侧发生故障时，系统不易失稳。”

马处长看着这样的结果，觉得所挖掘出来的规则还有一定道理，于是说：“这说明电力系统的暂态稳定问题实际上就是电网结构问题。若电网结构坚强有序，一般不会出现稳定事故；反之，事故难以避免。发生故障后，若线路切除使薄弱线路上的潮流降低很大，或不会使薄弱发电机与功率输送枢纽点间的电气距离因薄弱发电机与其余发电机间的电气距离增大而显著增大，则切除故障线路引起的网络结构改变会提高电力系统的暂态稳定性。”

下课铃响了，徐教授在屏幕上依次打出了自己的 E-mail 地址和电话：“大家谁有疑问可以通过邮件和电话跟我联系。”

3.4　负荷预测

今天的课在下午一、二节，徐教授走进教室发现有的学员还爬在桌子上，便说："孔子曰：'中午不睡，下午崩溃'。孟子曰：'孔子说的对'。看看大家的精气神儿，我就知道诸位都是中午睡了觉的。"

"徐老师真是太幽默了！"有学员喊道。

"这节课，看看数据挖掘的另外一个用武之地：电力负荷预测。"徐老师说出了本节课的主题。

"徐教授，我是搞客户关系管理的，电力行业真是门外汉。问个可能让大家见笑的问题：怎么理解电力负荷预测呀？"华润万家的万总谦虚地问道。

"问得好，这也正是我接下来要说的。电力负荷预测分长期、短期预测。实际上长期预测难度很高，它主要应用在电力规划、变电站的选址等，比较常见的是电力系统短期负荷预测。电力系统短期预测，顾名思义，是指预测未来一个月、一周或一天的电力负荷指标的预测。"徐老师简单地描述了一下。

"徐老师，那为什么要进行电力负荷预测呢？"南航的陆经理也踊跃地道出了自己心中的疑惑。

徐教授笑着说道："这个问题我看马处长比我有发言权，我们请电力公司的马处长给大家说说"。

马处长站起来说道："是这样的，电力负荷预测主要是为了电网供电容量的预安排。首先可以了解负荷时段与负荷量，从而安排电网内发电机检修及维护的顺序；其次，根据负荷预测安排电网建设计划，逐年投入新的机组，以满足根据预测出现的新负荷量。"

"马处长一定是经常接受记者采访。回答用两个字来形容是：完美，三个字来说就是：很完美！"，等下面的笑声小了点后，徐老师顿了顿，继续说道："有这样一

个故事：一个电工走入手术室，对一位戴着氧气罩的垂危病人说道：‘您好！请您深呼吸一次，这里需要停电五分钟！’一个简单的笑话，揭露出一个深刻的现实：电力负荷预测工作没做好，最后只能拉闸限电。”

“原来近几年的‘闹电荒’和电力负荷预测有关系”，航天研究院的黄主任若有所悟的感慨道。

“徐教授，各位老总，你们也都知道电力行业是垄断性质的，正所谓树大招风，一有差错，就遭话柄。因为这个拉闸限电，大家都很头疼。我们公司，甚至整个电网系统，都非常重视电力负荷的预测。只是缺少指导，不知道从哪里下手哇”刘总也坦诚给大家说出了心底的话。

徐教授便开始了专业知识的解说：“不要担心，数据挖掘来给你解惑。电力负荷预测有一大法宝：时间序列分析预测。时间序列是按照时间顺序排列的、随时间变化且相互关联的数据序列。分析时间序列的方法构成了数据挖掘的一个重要领域，即时间序列数据挖掘。要通过对时间序列的分析达到认识事物、了解其变化规律的目的，所用的方法主要是对给定的时间序列选择合适的数学模型。这个模型通常含有有限的未知参数，通过对这些参数的估计，最终建立起数学模型。当模型建立以后，就可以根据实际需要进行预报或控制。”

徐教授环视了一圈，看下面学员的反应，有的人在听，有的人貌似神游了。

徐教授为了活跃课堂气氛说：“听我讲完估计有人睡着了，除了签发‘特困证明’，我新想出一个主意，来帮助想睡觉的同学：就是我们悄悄换个教室，等那些睡觉的人醒来后就会发现，眼睛一闭一睁，老师和同学都不见了……”

专家一出手果真不同凡响，虽然下面没有学员真睡着，大家被徐教授的幽默逗乐了，瞌睡虫都被赶跑了。

徐教授接着说：“经过刚才的介绍，相信在座的各位已经明白了时间序列分析的基本思想。对时间序列分析的目的是找出数据的变化规律，即建立线性模型从而实现预测。时间序列研究的数据以一定时期（如一年、一月、一周等）为一周期呈现比较

有规律的上升、下降交替运动——即随着自然季节的更替发生有一定的规律性，比如淡季和旺季。”

刘总不淡定地说道：“这个正好和用户用电的特性吻合。按照周期为一年来说，也分淡季（一般是 4 月份左右）、旺季（一般是 8 月份左右）；按照周期为一礼拜、一天来说，也分用电低谷、高峰。比如居民用电，周末一般比工作日高，一天内晚上通常较白天用电高。”

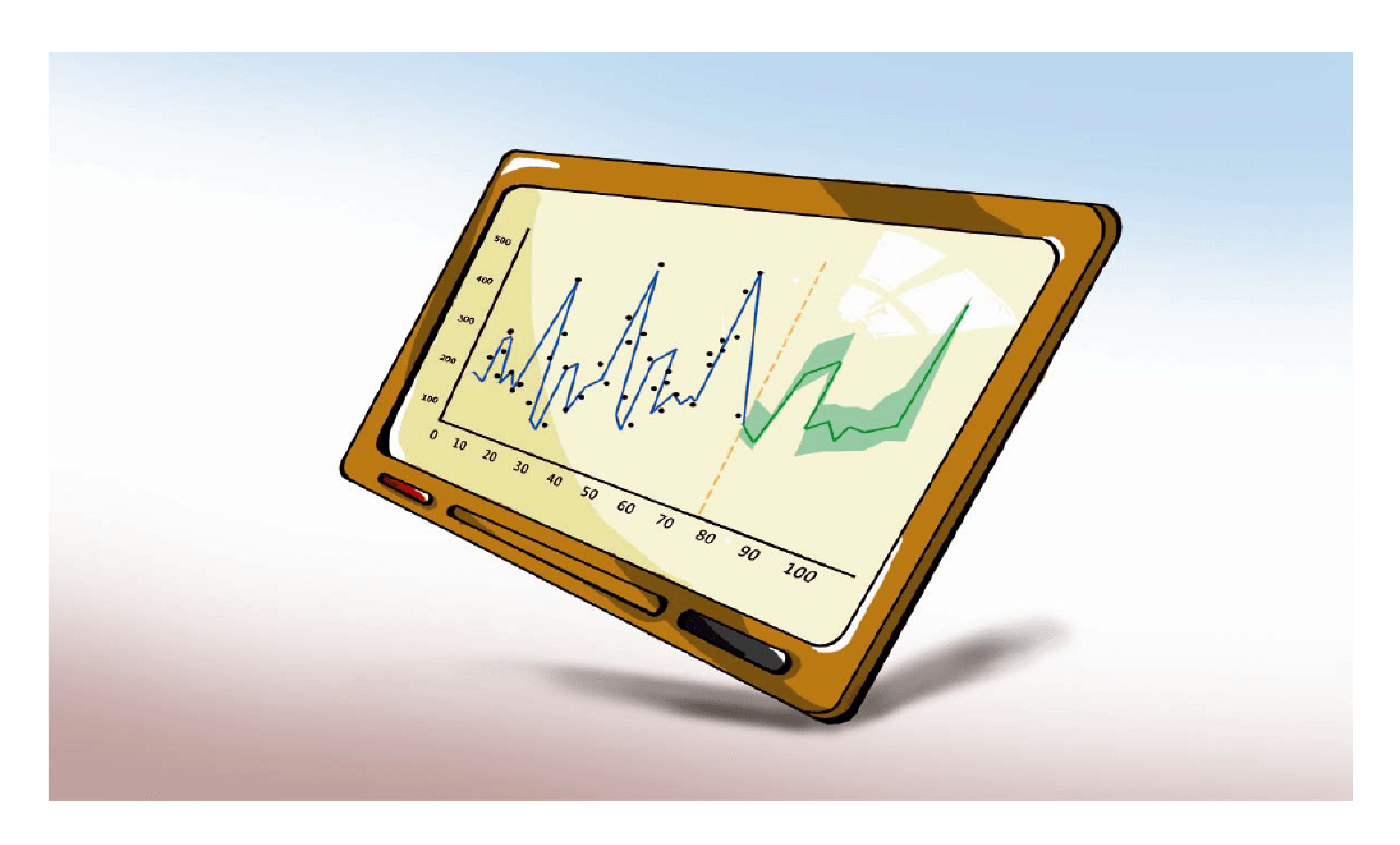
徐教授笑着回应说：“正是这样，所以时间序列数据挖掘方法可以用在电力负荷预测中。电力系统时间序列的建立首先要对样本数据进行分析并建立样本函数，然后依据单位时间内电力负荷用电量的样本函数而确立时间序列，最后进行预报。”

王总说：“我之前看过一个资料是关于电力负荷回归分析法的。回归分析电力负荷预测是通过对影响因子（如国民生产总值、工农业总产值、人口、气候等）和用电的历史资料进行统计分析，确定用电量和影响因子之间的函数关系，从而实现电力预测。”

徐教授说：“说得很好！在回归分析中，选用何种因子和该因子采用何种表达式只是一种推测，这影响了用电因子的多样性和某些因子的不可测性，使得回归分析在某些情况下受到限制。与回归分析的多影响因子分析不同，时间序列分析仅仅依靠过去某时间段的电力负荷值，建立模型后，直接预测未来的电力负荷。”

姚局长问：“徐教授，用时间序列分析后，预测的结果怎么样？”

徐教授解释道：“大家看图，竖直虚线右边曲线就是根据它之前的数据时间序列预测出来的。阴影区域是在一定的置信水平限制下，未来趋势可能出现的位置。”



台下一个学员说：“徐教授，这时间序列分析建模的过程想必很复杂。目前有没有什么比较成熟的方法？”

徐教授回答：“比较经典的建模方法有自回归模型、平均滑动模型、自回归滑动平均模型、求和自回归滑动模型等。前面四个模型针对的时间序列是平稳的。平稳时间序列认为序列其统计性质不会随着时间的推移而发生变化。这点要求是非常高的，实际中的大部分序列都是非平稳的。这时候，求和自回归滑动模型就可以帮助我们解决非平稳时间序列的建模问题。”

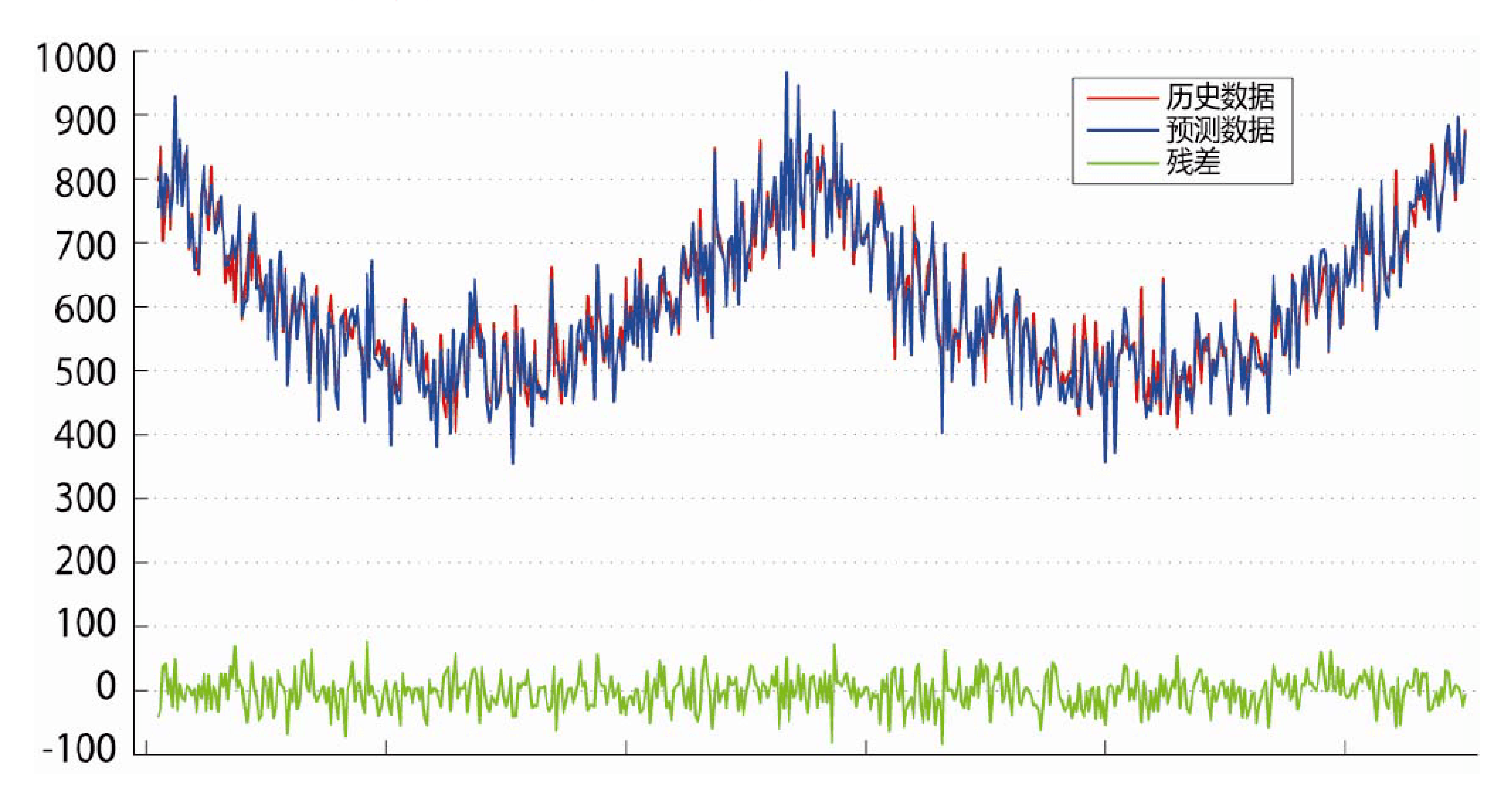
台下一个人感慨道：“听完这一堆名词，我头就大了。”

徐教授举例说：“以求和自回归滑动模型来说，第一步是进行模型的定阶和识别，即估计模型中的p和q，此步骤的标准有AIC准则（Akaike信息准则）等。在模型定阶和参数估计后，要对建立的模型进行考核，基本做法是检验模型的误差是否为白噪声。若是检验认为误差服从高斯分布，则建模获得通过。否则要重新进行定阶和参数估计。”

徐教授这番话之后，学员表示头大，专业知识太难理解。

移动的梁总分析说："对于电力系统负荷预测，曲线越接近目前的情况就应当越准确，而对于过去很久的数据，不必要作很精确地拟合，类似惯性作用。"

徐教授接着给补充道："梁总说得很正确。其实作为决策人，在座的诸位能理解宏观原理就可以了。外边盛传已久的'秘方'已经告诉诸位了，就是时间序列分析。秘方里面还有一个不可或缺的药引子，这里的药引子就是建模需要的数据。比如某人就是利用某地区5周的电力负荷数据，通过刚说的'秘方'，确定了'诊断良方'——预测模型，进而实现关于电力负荷的预报。"



"徐教授，那个诊断良方的效果怎么样啊？"刘总关切地问道。

"根据预测出来的结果，对比真实的电力负荷，误差是非常小（图中最下面的曲线）。且预测出来的结果很容易解释：周末由于工业负荷的减少，负荷水平普遍比平时下降。工作日内由于负荷受气温的影响较小，民用及工业负荷均较稳定。"徐教授边说边在大屏幕上展示模型的结果。

华润的万总说：“这个预测出来的精度很不错，大家看那个误差都在 0 的上下波动，就图中下面的曲线，完全处于可接受的范围内。”

“套用葛优代言神州行的话就是：数据挖掘，我看行！”，听完之后，刘总按捺不住内心的激动喜悦地说：“以后我们的电力服务水平有保障了，再也不用担心拉闸限电挨骂啦。”

3.5 盗电检测

EMBA 班的学员们都是各单位的领导，平时业务繁忙，没时间去享受大自然的旖旎风光。这次的数据挖掘课因徐教授出差改到了周六，在众学员的提议下，徐教授将本次的 EMBA 课堂搬出了室内，组织成为户外的爬山活动。



在爬山路上，李部长领头唱“红歌”，大家兴致勃勃地附和着唱，精彩程度真不逊色于全明星的“红歌演唱会。”在下山刚抵达山脚的时候，前面传来了吵吵闹闹的声音。大家凑过去一看，原来是电力稽查人员在现场抓住了两个实施窃电的贼娃子，双方正在进行着“拉锯战”，斗智斗勇……

在离开事发地段返回宾馆的路上，大家还在讨论刚看见的事情。“这窃电的家伙胆子也太大了，光天化日之下搞这样的勾当，真是吃豹子胆了！这不，被抓了个现行，不知错还想抵抗。”李部长感慨道。

马处长：“其实，盗电的行为还是比较多的，能抓住的很少，他们就更肆无忌惮了，我们电力公司每年因此损失高达 500 万元以上。”

吃过晚饭后，徐教授开始本次室外的课程，“今天下午大家看到有人光天化日之下盗电的情景，对我触动很大，刚才临时作了一个决定，将今天的学习内容改为基于数据挖掘的盗电检测方法。”

徐教授停顿了片刻，继续说：“如果你是电力公司领导，面对窃电行为的频频发生、窃电手段和方式的专业化、隐蔽化，该怎么办?”

“成立稽查大队，微服私访，实地侦察，必定会遏制这种现象的发生。”李部长率先开始献计献策。

“应该杀一儆百，发现窃电者，一律严惩不贷，这样有个威慑作用。”华润公司的万总说出了自己的看法。



S 钢铁公司的赵总也赶忙说出自己的意见："我建议建立远程集中抄表系统，实时监测用电情况，一旦有人窃电，立即发出报警信号。"

"窃电者玩高科技，以其人之道还治其人之身，我们也用高科技对付。就是不知道该用什么技术呀，得想想。"税务局姚局长喃喃自语。

徐教授点评道："综合来看，大家每个人都有一个侧重点，有从业务人员层面讲的，有从技术层面讲的，也有从经营管理层面讲的，点子都不错。我呢，其他也不懂，就知道一点儿数据挖掘，就从数据挖掘的角度说说吧。"

"太好了，徐教授，大家就等着数据挖掘这个神秘的利器出场呢，看看数据挖掘怎么成为火眼金睛的，能让这些电耗子原形毕露。"马处长说道。

“大家想一想，盗电用户的最大特点就是电费少交了，这与该企业的人员、产值、税收等形成反差，使其用电行为属性与正常的用户存在着很大差别。”徐教授分析道。

“徐教授，我的理解是，通过用电的消费行为差异性来区别普通用户和盗电用户，具体在技术上是通过什么方法来实现呢？”听完徐教授的介绍，马处长道出了心中的疑惑。

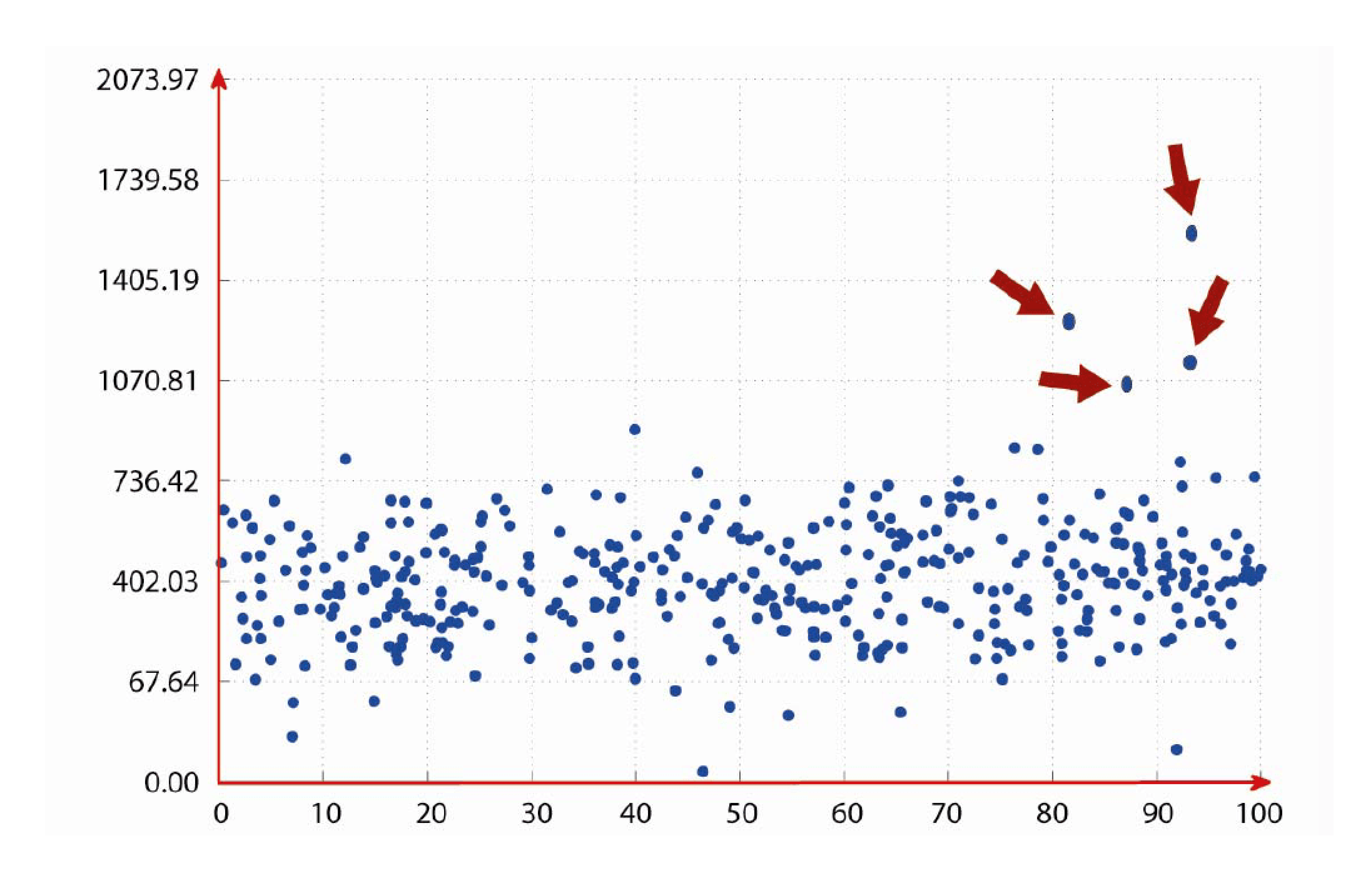
徐教授回答说：“你说得很对，盗电用户有着与普通用户不同的行为特征，必将成为孤立点。我们应用聚类分析方法，很容易让那些电耗子现出原形。”

马处长又问道：“徐老师，聚类就是一个‘类内相似性最大化，类间相似度最小化’的一个分群过程，聚类方法有很多种，上一周您讲过有基于距离的聚类，基于网格的聚类，基于密度的聚类，还有视觉聚类方法等，可我们到底使用那一种聚类方法进行盗电检测呢？”

“盗电检测就要进行孤立点分析，使用基于密度的聚类方法比较合适。”徐教授回答道。

李部长：“徐老师，以前您详细讲过基于距离的聚类方法，您再给我们描述一下基于密度的聚类方法的具体步骤吧。”

徐教授耐心地回答道：“基于密度的聚类方法主要包含以下几个步骤，（1）读入原始数据，并对这些数据进行（如缺失值、规范化等）预处理；（2）设定参数，即确定邻域半径大小或邻域内样本点最大数（即密度）；（3）确定邻域内对象，判断是否在邻域内，若不是继续选取样本点，若是最后输出结果。结果中，不包含在任何簇中的对象被认为是‘噪音’、‘孤立点’或‘异常值’，比如下图中的红箭头所示样本点，也就是可能的盗电用户。”



“通过对用户用电数据的聚类分析，反窃电的业务人员就能对锁定的目标重点侦查，可以有效地防止盗电现象发生。这样一方面提高了窃电客户识别率，同时还能节省电力部门人力资源，为反窃电工作提供了另外一种思路。”徐教授对本次的学习做了个小结。

税务局的姚局长于此得到启发：“哦，这么说除了电力行业的窃电检测，孤立点分析也可以用于银行的反洗钱侦察、税务部门的偷税、漏税活动甄别吧？”

徐教授肯定了姚局长的观点，并指出将在后续的课程中将给大家讲述这些内容。

3.6　电力数据挖掘系统的构建

一上课，徐教授说起数据挖掘在国家电网的应用动态：“熟悉电力部门的人都知道，今年国家电网成立了一个新部门：运营监测（控）中心，在总部和省公司两级部署。目标实现对公司经营管理 24 小时即时在线监测分析，实现对规划、建设、运行、

检修、营销、人资、财务、物资等业务全方位监测分析，实现对计划预算、资金收支、电力购销、资产全寿命周期、供电服务、产业发展、金融领域等全流程监测分析，构建集全面监测、运营分析、协调控制、全景展示于一体的综合管控平台。　”

航天研究院的黄主任说道：“徐教授，您这么一说，我也想起一个类似应用跟大家分享。我们曾经针对某研究建立过一个 IMS 系统，就是针对公司所有的 IT 信息系统进行监测和分析。包括桌面安全、业务系统使用情况的监控。我理解的是运营监测（控）中心是将对象扩大至整体业务的方方面面了吧。”

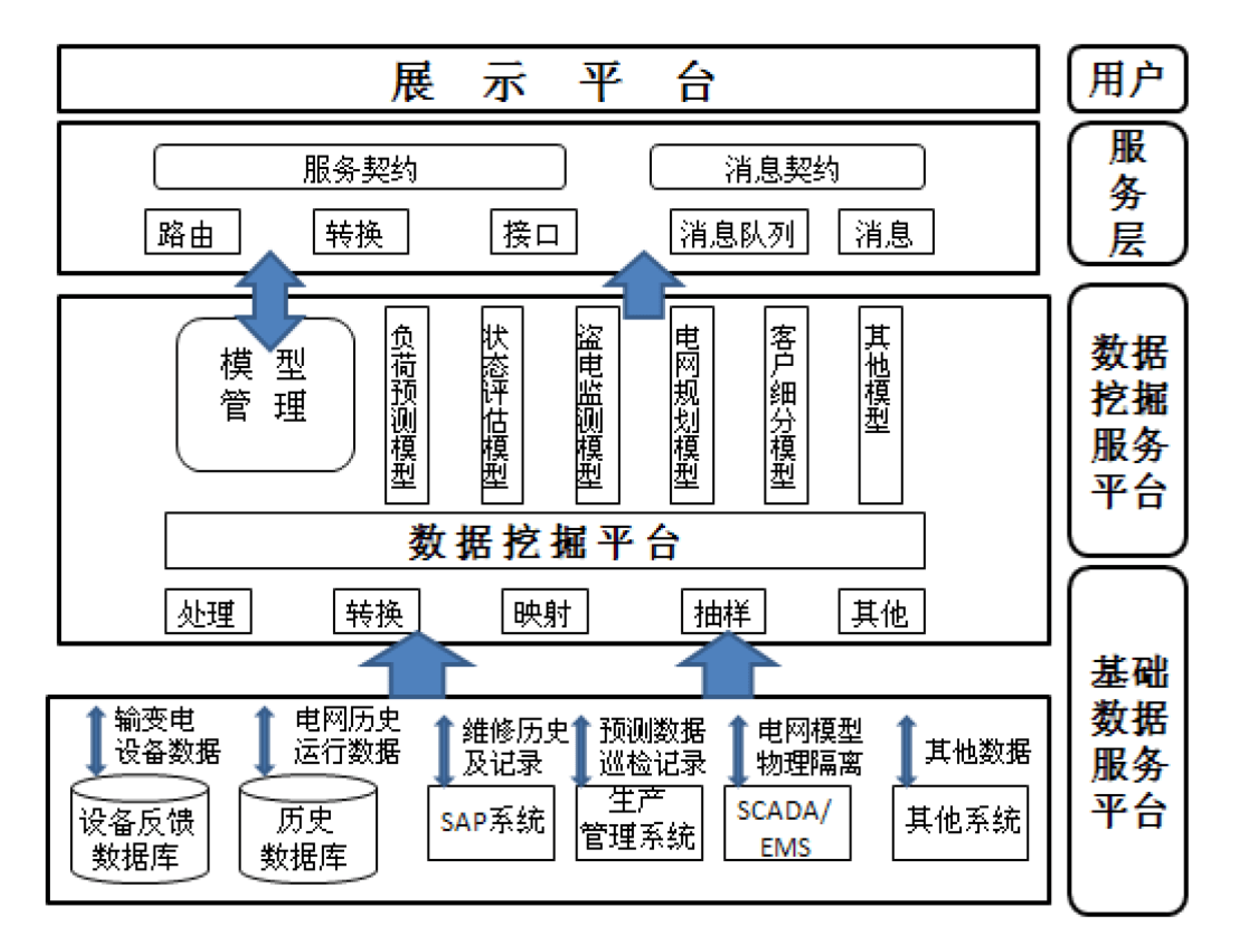
徐教授点头肯定道：“是的，你说的 IMS 和电网的运营综合管控平台，二者区别在于监测对象的差异。电网的监测和管控侧重点在全面性和重点业务的平衡把握，打破不同业务部门之间的壁垒，形成高效地协作机制。这就离不开数据挖掘，比如在监测内容的指标梳理、监测内容的高级深入分析上，数据挖掘都可以开展很多工作。利用数据挖掘技术建立公司综合绩效、发展能力、竞争能力、风险管控等方面的模型，对公司的整体运营情况中存在的异动和问题进行预警、分析，并协调解决。”

“徐老师，听了这几节课和您刚讲述的电网运营监控，我感觉数据挖掘确实能够在我们电力行业有广阔的应用前景，很有必要构建电力数据挖掘平台，可怎么建立呢？”马处长急切地问。

“若想构建企业级数据挖掘系统，最好先建立企业级的数据仓库。”徐教授建议说。

“数据仓库？好办，我们已经花了两年时间建立了电力数据中心，在此基础上，考虑到数据挖掘的各种主题，如设备状态检修及寿命评估、电力稳定性分析、负荷预测、盗电检测和规划设计等，很快就会建立起支撑数据挖掘的电力数据仓库。”马处长激动地说。

徐教授说：“好，下面给大家简单介绍一下企业级数据挖掘平台的体系结构，请大家看大屏幕。”



“处于最底层的是数据服务层，对来源于异种结构的数据进行转换、映射、清洗等操作，为数据挖掘进行数据准备。数据挖掘服务平台主要用来实现各种模型的建立。服务层是展示平台和数据挖掘服务平台的中间纽带，管理和控制各专业模块，并建立与数据库的连接，响应用户的操作请求。”

马处长激动得不得了：“不错！不错！如果有了电力数据挖掘平台，我这个副处长就可以官复原职啦！”

教室里一阵笑声。

第 4 章　数据挖掘在交通航空领域的应用

1978 年，邓小平造访日本。期间，日方安排邓小平搭乘世界首条载客营运的高铁——日本新干线。坐在宽敞豁亮的车厢中，凝视着窗外急速飞过的模糊风景，邓小平感慨称：“新干线推着人们跑，我们现在很需要跑。”斗转星移，随着国内高速铁路的快速发展，“中国高铁”的梦想（像风一样快）已经成为现实。与此同时，高铁也引起了人们的广泛关注，其票价和安全性等无一不是民众时常议论的话题。

此外，智能交通系统是近年来迅速发展的城市道路、高速公路控制管理的新技术。该系统是由先进的交通管理、控制、营运调度等信息系统组成。其目标是将运输系统中的人、车、路三要素紧密地结合在一起，最大限度地发挥整个交通系统的效率。良好的交通流量预测，是智能交通系统的实时交通信号控制、交通分配、路径诱导、自动导航，事故检测等的前提。

鉴于此，徐教授专门安排两节课来讨论数据挖掘在高铁票价的制定、高铁轨道安全性检测和交通流量预测中的应用。

4.1 铁路票价制定

徐教授开课讲道："近年来，我国高速铁路建设非常迅速。根据铁道部的规划，到 2020 年，全国将建设高速铁路 1.6 万公里以上，铁路快速客运网将覆盖全国 90%以上人口，形成'四纵四横'的高速铁路网。"

谈到这几年中国高铁的成就，铁路局高局长喜形于色，骄傲地说："中国高铁在短时间内密集地取得了一系列成果：2008 年 8 月 1 日，中国第一条具有完全自主知识产权、世界一流水平的高速铁路京津城际铁路通车运营，最高运行时速 350 公里。2009 年 12 月 26 日，世界上里程最长、工程类型最复杂的武广高速铁路开通运营，创造了时速 350 公里隧道内会车、两列重联条件下双弓受流等一系列世界新纪录。武广高铁昭示着我国能够建设工程类型齐全、大规模、长距离、世界一流的高速铁路。2010 年 2 月 6 日，世界首条修建在湿陷性黄土地区，时速 350 公里的郑西高速铁路开通运营，标志着我国能够在国外未曾预见到的特殊复杂地质条件下建设世界一流高速铁路。2010 年 7 月 1 日，沪宁城际高速铁路的开通运营，是在深厚软土地区建设速度最快、运行速度最高的高速铁路。2011 年 6 月 30 日京沪高铁正式开通运营。作为新中国成立以来建设里程最长、投资最大、标准最高的高速铁路，京沪高铁贯通'三市四省'，串起京沪'经济走廊'。"

听着听着，马处长轻轻地敲了几下桌子："高局长，你说的全国人民都知道，成绩不说跑不了，问题不说不得了！"

高局长被马处长的当头一棒打懵了，还没来得及反应，李部长也开始抱怨起来："前一段时间，网上盛传的一张沪杭高铁'一人一车厢'的照片，局长大人难道视而不见吗？"



“您再听听老百姓的呼声吧！”航天集团的黄主任说话时情绪有些激动。

学员们将街上听到的、网上和电视上看到的人们对高铁票价的怨恨纷纷暴露出来：

“太贵了，高铁票价是普通绿皮车的 7～8 倍……”

“买不起，不考虑坐。高铁，价格那么高，乘客怎么会去追捧？够不着！”

“高铁是贵族专列，啥时候能照顾一下小老百姓呀！”

“用大部分纳税人的钱来建设，却只让小部分人乘坐，浪费国家资源！”

“……”

徐教授打了个停止的手势，说道：“对于高铁票价问题，我们听听铁路局高局长的介绍，一起来了解一下铁路部门是怎么应对的。”

经过刚才的“炮轰”，高局长这次谦卑多了，站起来答道：“高铁以 350 公里运行的时候，武广高铁上座率比较低，不足四成。应对大众高票价的质疑声，武广高铁采取的主要手段是通过降速来降低票价。现在高铁的运行速度都已经降速至 300 公里以内，上座率也提高到 74%。”

马处长问道：“高局长，目前的武广高铁票价是个什么水平？购高铁票时有什么优惠政策？”

高局长回答说：“目前武广高铁分一等高铁票和二等高铁票，每公里价格分别为 0.729 元和 0.459 元，无打折计划。普通人购票时无个人优惠，但是有团体优惠。在非春运期间，满 20 人团购可免收 1 人票价，20 人以上每增加 10 人再免收 1 人票价。”



李部长经常全国各地跑，对于出行选择乘飞机还是高铁，他颇有研究："其实高铁的二等票相当于飞机经济舱的 5～6 折之间。因为飞机要有机场建设费和燃油费，所以一般情况下 5 折以下的机票在价格上可以和高铁比拼。"

南航的陆经理表示："是的，这给我们民航部门的飞机运行带来了很大压力。有专家预测：在中国高铁高速发展的当下，如果高铁能够适当降低票价，就能吸引更多乘客，从而形成规模效应，实现良性循环。到时候，我们民航真的要喝西北风了。"

高局长也为难地说："都一样，我们高铁部门的日子也不好过啊，现在国家铁路整体负债率高达 60%，直逼国际负债警戒值。大家都知道高铁运营成本高，如果票价太低就很难偿还债务，更别说盈利了。"

徐教授："总体来说，高铁定价是个很复杂的问题，要顾及多方面的因素。比如老百姓的消费水平，铁路部门的成本回收问题，还有民航等竞争部门的利益等。但是，目前国内的高铁票价制定还基本上是停留在根据线路运营里程乘以单价的方式上。"

马处长又提问道："以路线长度乘以单价来计算高铁价格，这个方式的好处是比较容易理解，便于业务人员管理。徐老师，您觉得目前国内的这种高铁定价方法有哪些不足之处？"



徐教授说道："从票价来说，目前划分是一等票价和二等票价。即使是减速降价，高铁的票价还是过于'一刀切'。若是淡季、旺季，早班、晚班，直达、停站多的列车班次全都采取统一票价，这种票价体系还是不够灵活。"

李部长也问道："高局长，在高铁票价制定上，国外有什么先进经验值得我们借鉴？"

高部长对此曾经花时间研究过，畅谈起来："世界上大多数国家的高速铁路，都采取丰富的差异化定价。比如德国提供了复杂的价格优惠制度：不经常乘坐火车的旅客往往都会购买一张火车票打折的年卡。50 欧元的年卡在全年任何时候购票都可以享受 7.5 折优惠，200 欧元的年卡则可以享受 5 折优惠。另外，买往返票会有折扣；如果往返行程中，隔着一个周末，又有优惠；提前 24 小时、72 小时、7 天、14 天购票的优惠幅度是不同的。"

鼓风动力集团的王总接着问道：“未来，竞争将迫使我国铁路部门重新考虑以优质服务吸引客源，比如降低非高峰时段的票价。世界上像德国这样实行高铁票价优惠的国家多吗？”

高局长继续回答道：“基本上发达国家高铁定价都有优惠。比如法国，有一种国家规定的优惠政策，主要内容有家庭成员外出坐火车，三个以上小孩，最少可以减价 30%、最多可以减价 70%。另外，每天工作往返同一条线路优惠，对军人乘车可以优惠，军人自己出的票价 25%，铁路部门出 24.6%，国家补助 50.4%。”

徐教授总结道：“虽然各个国家采取的优惠措施有所差异，但彰显的是一个事实：合适的定价是提高高铁上座率的保障。”

高局长：“徐老师，那从数据挖掘的角度来看，能为高铁定价提供哪些思路，您给我们支支招吧。”

徐教授回答道：“第一个典型的手段就是通过聚类分析将市场切割为不同的市场，根据旅客消费特征的差异性确定不同价格。这个在实际应用中也很容易理解，因为不同群体的消费意愿和支付能力不同，因此需求价格的弹性也有高有低。这样针对不同细分市场制定不同价格，采取各种营销活动，就可以实现利润最大化的目标。”

高局长对徐教授讲的更加感兴趣了，继续请求道：“徐老师，您还是讲一个实际的例子，让我们更容易地理解这种市场细分的手段吧。”

徐教授：“大家知道，法国的高速公路非常发达，承担了法国国内 90%的客运量和 60%的货运量，所以其铁路部门的竞争压力非常之大。2000 年以前，法国铁路部门每年都在亏损，政府每年都要补贴铁路部门几十亿法郎。”

马处长诧异地说：“铁路部门为垄断行业，还赔钱，不可思议！”

高局长也笑着说：“是的，确实是这样。”

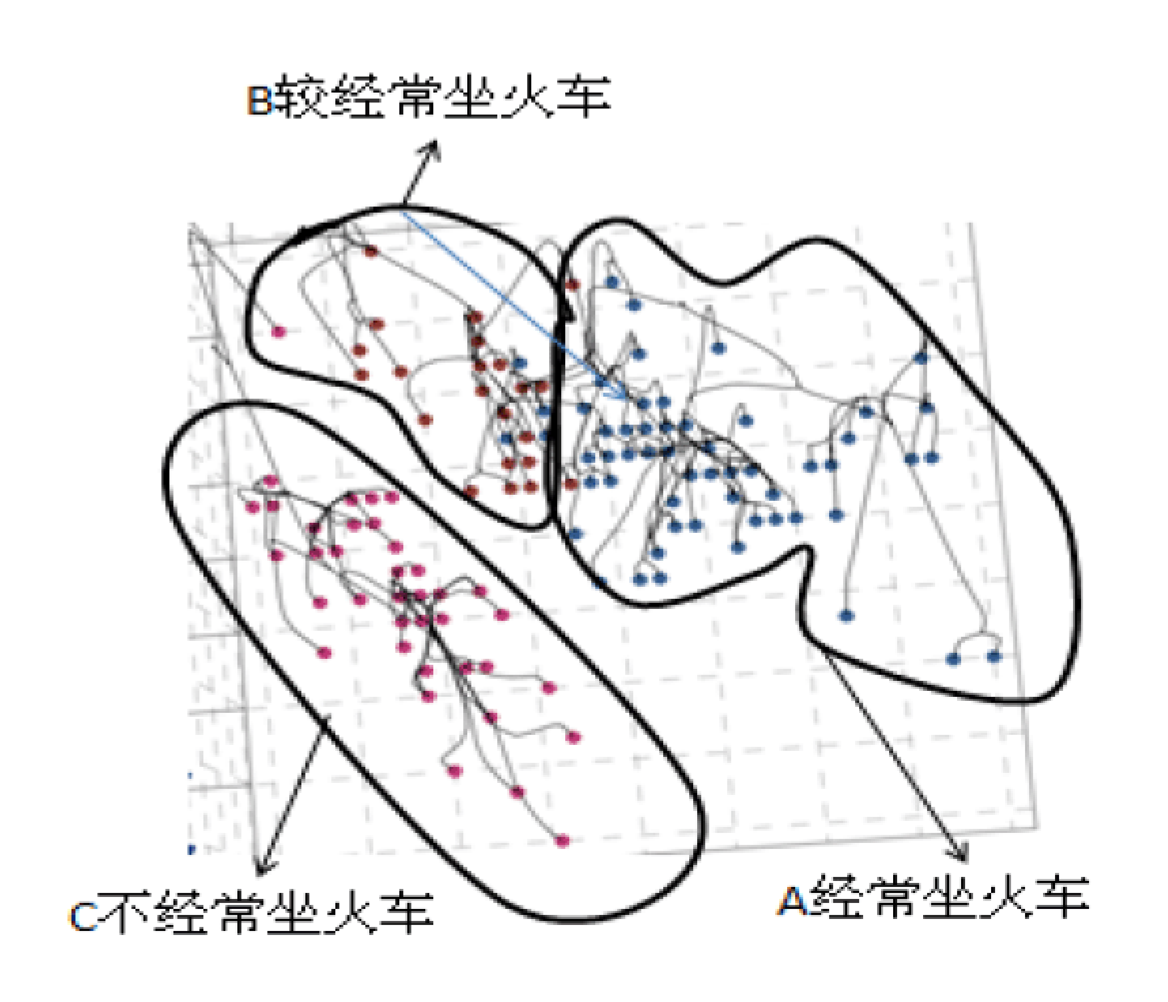
徐教授：“不过，后来局势被扭转了，秘密武器就是数据挖掘。”

高局长兴趣更浓了："原来我并不知道其中玄机，徐老师您就带领我们开开眼界吧。"

徐教授继续刚才的话题说道："法国铁路部门在不断提高服务质量的同时，制定高铁票价时，尊重价值规律，在不突破国家定价的基础上，依据旅客市场细分结果来制定高铁票价。"

高局长对聚类算法已经比较了解了，他觉得对旅客数据进行聚类分析已胸有成竹，便说道："相信我们大家现在都不会关注聚类的过程了。徐教授，你还是详细地给大家讲一讲聚类后对不同旅客群体如何设计营销策略吧。"

徐教授调出一张 PPT 来帮助学员理解："聚类的结果显示，旅客可分为三类。经常坐火车的 A 类：优惠的幅度最大，积累一定的里程数，可免费乘坐，而且可以享受其他的优惠待遇。比较经常坐火车的 B 类：可以花一定费用来买优惠卡，最大减价幅度可达 50%，使他们由比较爱坐火车过渡到经常坐火车。偶尔乘坐火车的 C 类：主要是 60 岁以上老人、25 岁以下的年轻人，对他们减价 25%，使他们对坐火车逐步感兴趣。"



徐教授将光笔指向 PPT 图上尖尖重叠的部分，继续说道："观察聚类的动态过程，可以发现 A 类立体簇和 B 类立体簇有一定的交叉部分，这说明部分旅客同时满足优

惠 A、B 条件，只要实际中限制不可重复享受优惠即可。这是由于在聚类过程中的坐车频次相关参数设置区域交叠引起的。”

高局长说道：“明白了，应该是按照乘车频次初步分了三类旅客：不经常坐火车的、比较经常坐火车的、经常坐火车的。然后根据年龄、乘客身份、乘车档次等信息进行了群体聚类再划分。最后针对这些群体，制定并实行差异化的服务策略。”

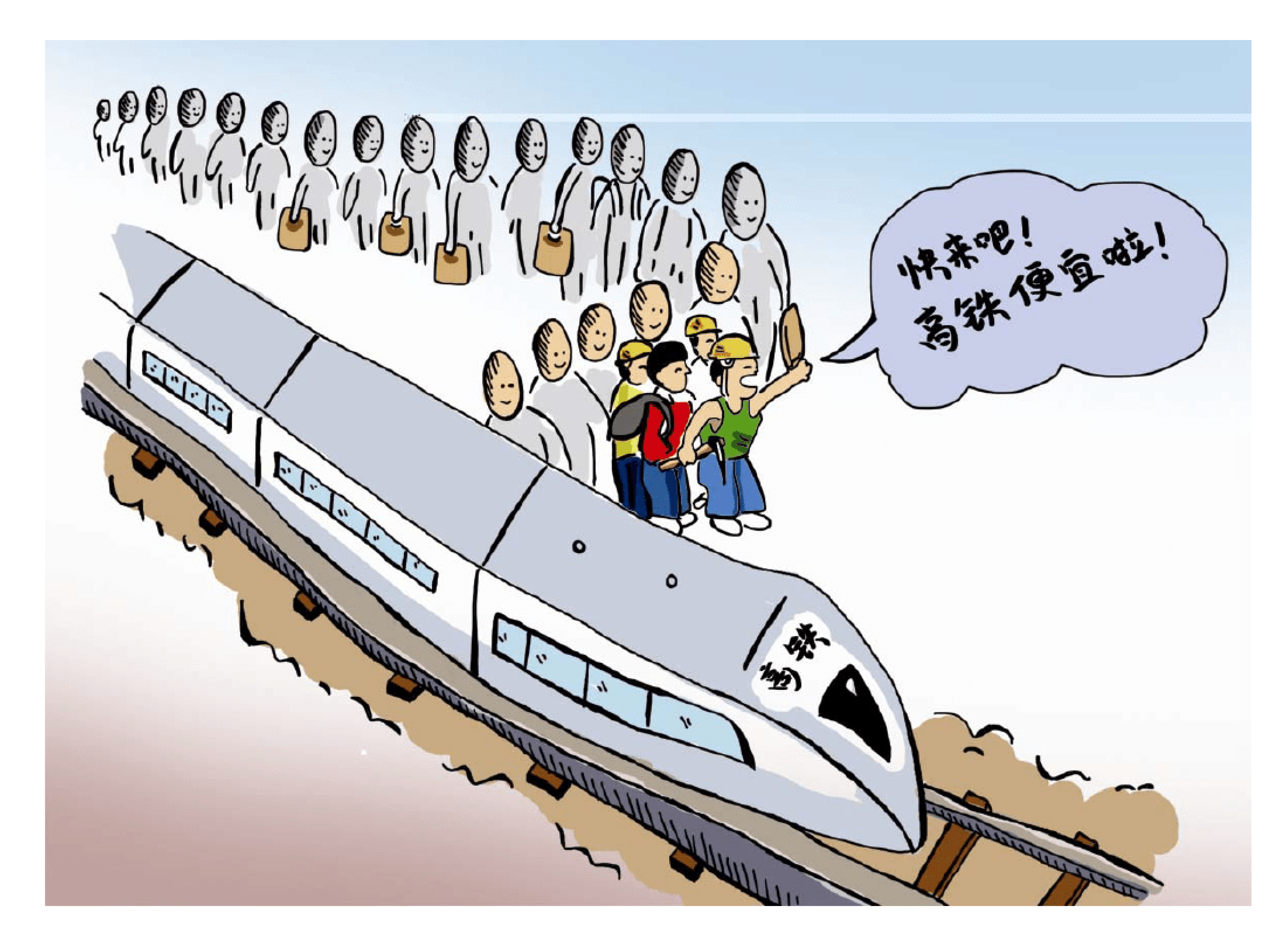
“是的，理解地非常到位。当然减价原则并非一成不变的，还要根据具体的情况确定具体的减价幅度。如旅游淡旺季的变化，运行高峰、低谷的变化，铁路运行线路是否有竞争对手等情况确定价格优惠的幅度，”徐教授进一步说道。

高局长说道：“顾客差异化定价就是企业对同一产品，根据不同的销售对象、不同的消费地点和不同的销售时间、不同产品等方面的需求差异而制定不同的价格。随着社会阶层分化，公众的个性化需求不断增加，这样的好处就是采用服务和价格的不同搭配销售策略，向用户提供更多选择。除了这个细分市场，在票价制定上还有什么数据挖掘方法可以指导票价制定，从而实现经营利润最大化呢？”

徐教授继续说道：“票价制定的另外一种方法就是通过回归分析等技术手段对高铁票价进行动态预测。”

高局长听到徐教授还有一招，真是喜出望外，急忙问道：“徐老师，高铁票价预测需要考虑哪些变量呢？”

徐教授解答道：“一般需要考虑以下几种因素：（1）高铁运行成本，包括运行距离、运行时间、沿线路网耗费、旅客流量等；（2）竞争对手的价格，如航空、公路的优惠情况；（3）市场的周期性，比如淡、旺季、节假日等信息。不同预测计算中，上述几种因素考察的重点应有所不同。以运行成本为例，可分为平均成本定价法、盈亏平衡定价法、目标收益定价法、变动成本定价法、边际成本定价法等。因为市场中高铁票价也是一个动态发展过程，所以预测出其未来一段时间的变化趋势，实行浮动式的高铁票价制定，必将有助于灵活应对市场不断变化的需求。”



高局长得到了启发，也认同地表示："根据分析结果，可以策划各种促销活动，例如根据不同时段提供的不同折扣，往返票、多次票或者月票、年票都应该有不同的折扣，就能最大程度地刺激顾客的车票购买欲望。"

徐教授建议道："还可以实行会员制，经常乘坐高铁的旅客可以享受优惠价、积分换里程，或者免费车厢升级服务，从而满足不同层次群体的需求。"

下课铃响了，高局长提议说："让我们以热烈的掌声感谢徐教授给我们带来如此精彩的关于高铁票价制定的技术方法。但愿我们铁路部门能够采用这些先进技术，给老百姓带来真正的实惠，使全国人民人人都能享受中国高铁的丰硕成果！"

徐教授带头拍手，教室里掌声震天，大家都对未来高铁票的合理定价充满信心。

4.2　高铁轨道检修

“台下的都是成功人士，肯定为飞机和高铁贡献了不少 Money！我们先来做个小调查：不知道大家选择出行工具时，是选择灰机还是高铁？”徐教授开场说道。

“徐老师，尽管现在飞机都成灰机了，我还是坚定不移地选灰机。”李部长带头给出了自己的选择。

“高铁吧，与时俱进嘛。灰机航班经常延误。”马处长也跟着说出自己的抉择。

经常出差的万总说道：“乘高铁吧，飞机不安全，总觉得离地了就没安全感。听说葛优就是因为恐高从来不搭飞机！”

南航的陆经理也表达了自己的观点：“高铁也不见得安全，2011 年 7 月 23 日晚上 20 点 30 分左右，甬温线永嘉站至温州南站间，北京南至福州 D301 次列车与杭州至福州南 D3115 次列车发生追尾事故。所以我还是选乘我们民航的飞机，技术上相对新兴的高铁更成熟一些。”



“是啊，国内高铁时速都 380 公里了，这速度快了也让人担心。之前听朋友调侃高铁：‘按中央气象台的路径显示，10 年来威力最强的台风梅花会擦着南京过啊！干脆把我吹回北京得了，估计比坐高铁安全……’”S 钢铁公司的赵总道出了人们对高铁安全的担心。

电信公司的冯总也打趣说：“是啊，现在高铁安全已经被提升至风口浪尖了。前些日子，就有全副武装的高铁安全帽哥引发网友围观。他所必备的乘车设备有安全帽、自制安全带、手电筒、瑞士军刀、扇子、花露水、DV、雨伞、云南白药……真可谓是全副武装！”

“大家说了这么多，实际上可以一言而蔽之：高铁的安全性问题。”徐教授顺势引出这一节课的内容。

高局长听到徐教授又要讲解高铁安全问题，高兴得合不拢嘴：“运营安全是高铁的核心，高于一切。高铁安全是靠系统工程来保障的，整个高速铁路的建设过程，无论是从勘察设计、建筑工程、产品设备安装工程，都严格地进行质量控制。高铁列车运行的机械化和自动化程度非常高，有着极高的安全系数。”

听着高局长的自我陶醉，马处长有点不淡定了：“都追尾了，还自吹自擂什么！”

徐教授赶紧灭火：“我国的高铁经过十几年的引进、消化吸收、创新，谦虚地说，从总体上已经赶上了国际先进水平。不谦虚地说，在很多方面已经代表着国际水平。”

徐教授故意停顿下来，观察了一下大家对他所说的话的反映，见没有人异议，于是继续说道：“看来大家还是认可我的说法，下面我们以高铁轨道检测为例，说一说数据挖掘技术在高铁的安全保障中的应用。”

税务的赵局长说起自己曾经奔赴日本考察，在乘新干线火车时了解到的情况：“据同行车上的轨道维修师傅描述，他们是根据轨道轨检车每 10d 检测一遍的具体资料确定工作量。由于新干线为客运专线，轴重轻（原来轴重为 15 吨，现减小到 11 吨），板式轨道比重大（板式轨道占 53%），故轨道几何尺寸变化较小。”

高局长对徐教授的上课模式已经摸透了，不用说，徐教授肯定让他介绍高铁轨道检测现状。果然徐教授抬手示意他表达高铁现状，于是他站立起来说道：“随着高铁的繁忙运行，日客流量的不断增长，高铁线路的几何形位也会产生变化，轨道结构也会产生损坏。现在多采用雷达检测车对路基和轨道进行快速、无损地连续检测，还可以进行多通道载荷、位移测试。其检测速度一般能够达到 40km/h，而且排除了各种人为因素的干扰，获得真实信息，这就为轨道设备保持良好状态提供了保障。”

徐教授补充道：“在检测过程中，收集了图像传感器、超声波探伤、雷达测试、激光光电、其他各种轨道检测设备获取的大量信息，比如仅 1 公里线路上轨检车最高采样点可达到 4100 多个。”

序号	轨道质量相关数据
1	轨道几何尺寸（轨距、水平、轨向、尖趾距离、查照间隔）
2	钢轨（接头轨面、内侧错牙、轨端飞边、轨缝）
3	轨枕（碎石道床轨接头岔枕、整体道床轨枕玻璃钢套管）
4	联接零件（尖轨、可动心轨与滑床板间缝，弹条中部前端下颏离缝）
5	轨道加强设备（转辙、辙叉部分轨撑离缝，爬行量）
…	……

“徐老师，有了这些数据就可以进行数据挖掘了吧。”高局长问道。

徐教授回答道：“是的。对轨道的轨距、方向、高低、水平以及曲线超高、曲率、车体的水平振动加速度、车体垂直振动加速度等历史数据进行清洗、归一化处理后，通过回归等数据挖掘手段建立轨道状态检测模型。根据德国高速铁路的实践经验，直接影响及控制行车速度的主要因素有两个：一是轨道线路平纵断面，另一个是轨道线路的平顺性。所以在建立轨道状态检修模型时，着重选取对线路平纵断面和平顺性的指标，如横向震动加速度、轨向、高低等数据。”

高局长更具体地问道：“应用数据挖掘技术，还能够建立哪些回归模型？”

徐教授继续回答道：“主要是分析和研究轨道的动力学特性：（1）高速铁路轨道在动荷载作用下的特性和规律；（2）轮轨接触不平顺作用下的垂向受力与变形关系；（3）高速动载作用下刚度与阻尼对轨道性能的关系；（4）高速动载作用下列车临界速度和路基状况对无砟轨道性能的影响等。”

高局长进一步问道：“徐老师，这些回归模型确定后，怎么应用呢？”

徐教授解释道：“有了这些模型，再利用可视化技术可以监测轨道状况。比如，对轨道部件状态可以全面监测，如扣件脱落、螺栓松动、鱼尾板断裂、钢轨磨耗、道床路基水浸、坍落等异常状况。”

高局长似乎还不放心，又问道：“徐老师，利用数据挖掘进行轨道检测，有没有实际的应用？”

徐教授肯定地回答："某高铁公司在 2005 年向美国 ENSCO 公司定购的大型轨道检查车，通过建立的模型，将各种大型检测机械的检测数据代入后，真实地反映轨道的实际状况，在正挂、反挂、顺跑、逆跑都不会产生方向性问题。"

华润的万总更关注策略的落地，问道："这个轨道状态检测模型能不能直接指导工作人员的具体操作呢？"

徐教授解释说："以法国高速铁路为例，它通过车体振动加速度和转向架振动加速度来评价轨道质量状态。其轨道状态检测模型最终按轨道的质量状态分为四级：

轨道质量状态	对应质量状态描述
目标值（VO）	新线铺设、维修作业后应达到的质量标准
警告值（VA）	达到或超过此值的轨道不平顺，要实施重点观测，分析其发展变化情况并做出维修计划
干预值（VI）	达到或超过此值的地点或区段要实施必要的维修作业，一般在 15 天之内予以实施，并使其达到目标值
限速值（VR）	达到或超过该值的地点或区段列车必须降速行驶，并以任何可能的手段包括手工作业予以整治

有了轨道质量状态的评价和控制，业务人员就可以很方便地按照指示完成任务了。"

高局长总结道："轨道检修的目的就是为了确保高铁运营的安全性，尽可能地延长车辆的使用寿命，从而降低高铁的运营成本，提高效益。通过数据挖掘技术研究出的轨道养护维修模式，可以帮助我们更好地理解状态检修的内涵，必能为我们轨道交通养护维修提供借鉴、指导。"

4.3 交通流量预测

"在这一节课开始之前，我先给大家讲一个笑话。"徐教授看着大家说道。

"某君带着一只乌龟，下班后开车回家，在二环路上遭遇堵车。看汽车半天走不了几步，乌龟耐不住性子，坚持要先爬回家去，主人只好由它去了。不知过了多久，

主人在车里听见敲门声，打开一看，只见乌龟满头冒汗，气鼓鼓地说：‘你忘了给我家门钥匙……’。”

大家联系起自己堵车的经历，都忍不住笑了。



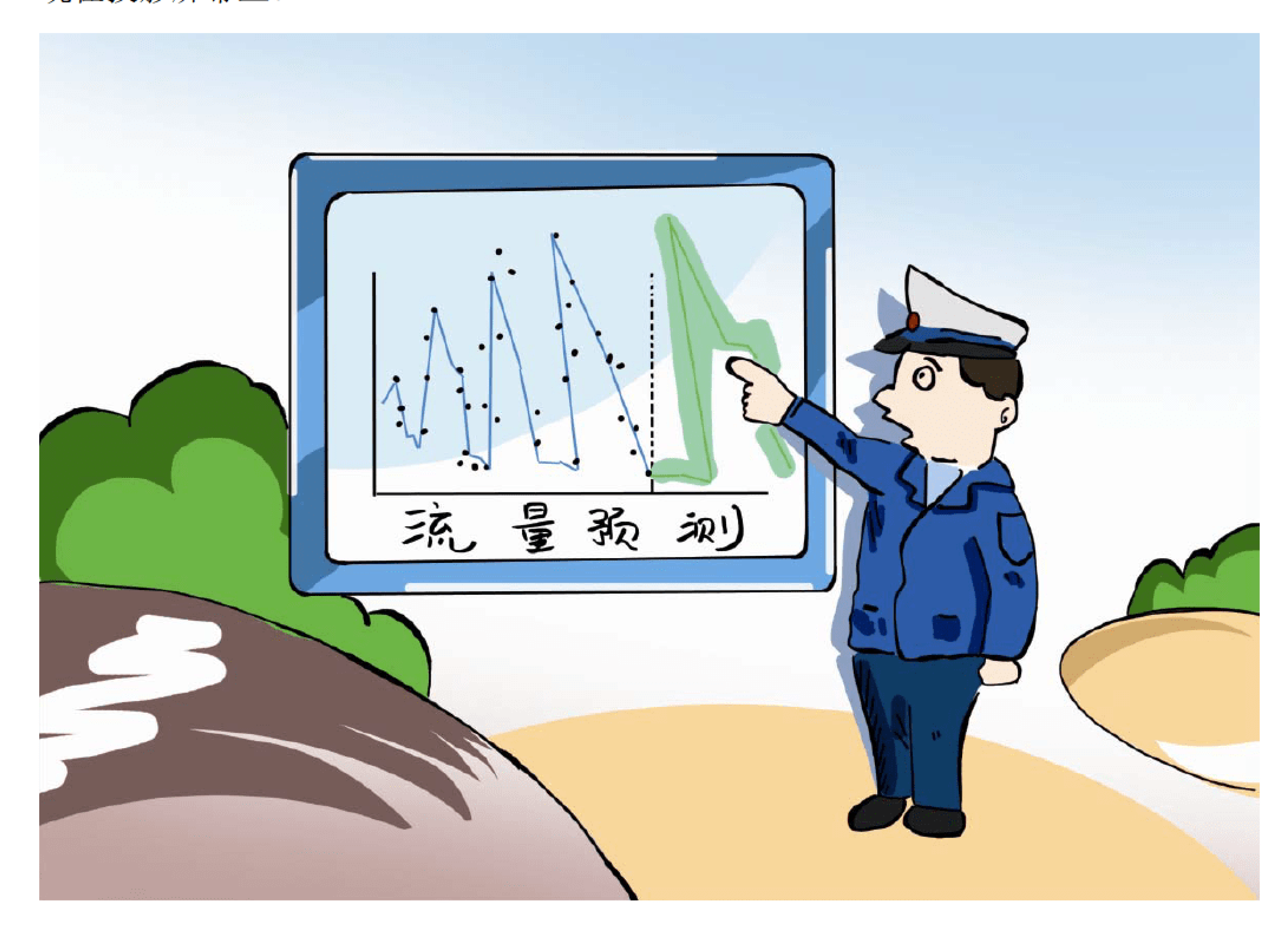
徐教授示意大家静一静，接着说：“今天这节课的主题就是运用数据挖掘技术进行交通流的预测，为交通调度策略制定、道路建设和改造提供决策支持，从而避免堵车发生或者减少堵车时间。”

徐教授招手让坐在最后一排穿制服的交警到前排来，并介绍说：“这位是我们学校隔壁交警一大队的刘队长，他听说我们今天要讲交通流量预测，特意前来听课。首先由刘队长给大家介绍一下交通流预测问题及其现状。”

刘队长迈着正规的步伐走上讲台，给大家行了个警礼后，开始讲到：“大家好！很荣幸有机会和大家一起聆听徐教授的数据挖掘在交通流预测中的应用这节课。”

停了几十秒钟，刘队长接着说："道路交通系统是一个有人参与的、时时变化的、复杂的非线性系统，交通流量除了受一些周期性的因素如节假日、季节影响之外，还具有很多不确定因素，如路面状况、天气变化、突发事件等，这些因素都给交通流量预测带来了一定的难度，特别是短时交通流量预测更加困难。"

说到这儿，刘队长熟练地将徐教授的笔记本电脑以自己的 3G 智能手机为路由连上了 Internet，然后进入到交警的城市交通指挥网。几个十字路口川流不息的景象展现在投影屏幕上。



刘队长指着屏幕说道："现在，我们每时每刻都可以通过智能交通控制系统 SCOOT 获取大量的交通信息，如来自分布于主要交通路口和干道的 360 多个摄像机的视频信息，来自 1300 多个传感器数据、道路占有信息、来自车辆定位系统的行程时间、平均速度等信息，每个月所产生的数据量达到上百 GB。"

徐教授趁机说道：“好啊，有了大量的实时数据，我们就可以利用数据挖掘进行交通流量预测了！”

刘队长想起了半年前他与徐教授所讨论的问题，说：“徐教授，就像您曾经给我们分析过那样，交通流具有不同的空间分布模式，例如城市主干道的交通流具有‘线’性模式，交叉路口的交通流具有‘平面’模式等，对城市道路交通网络进行实时、动态的交通区域划分是当前智能交通系统的研究难点之一。”

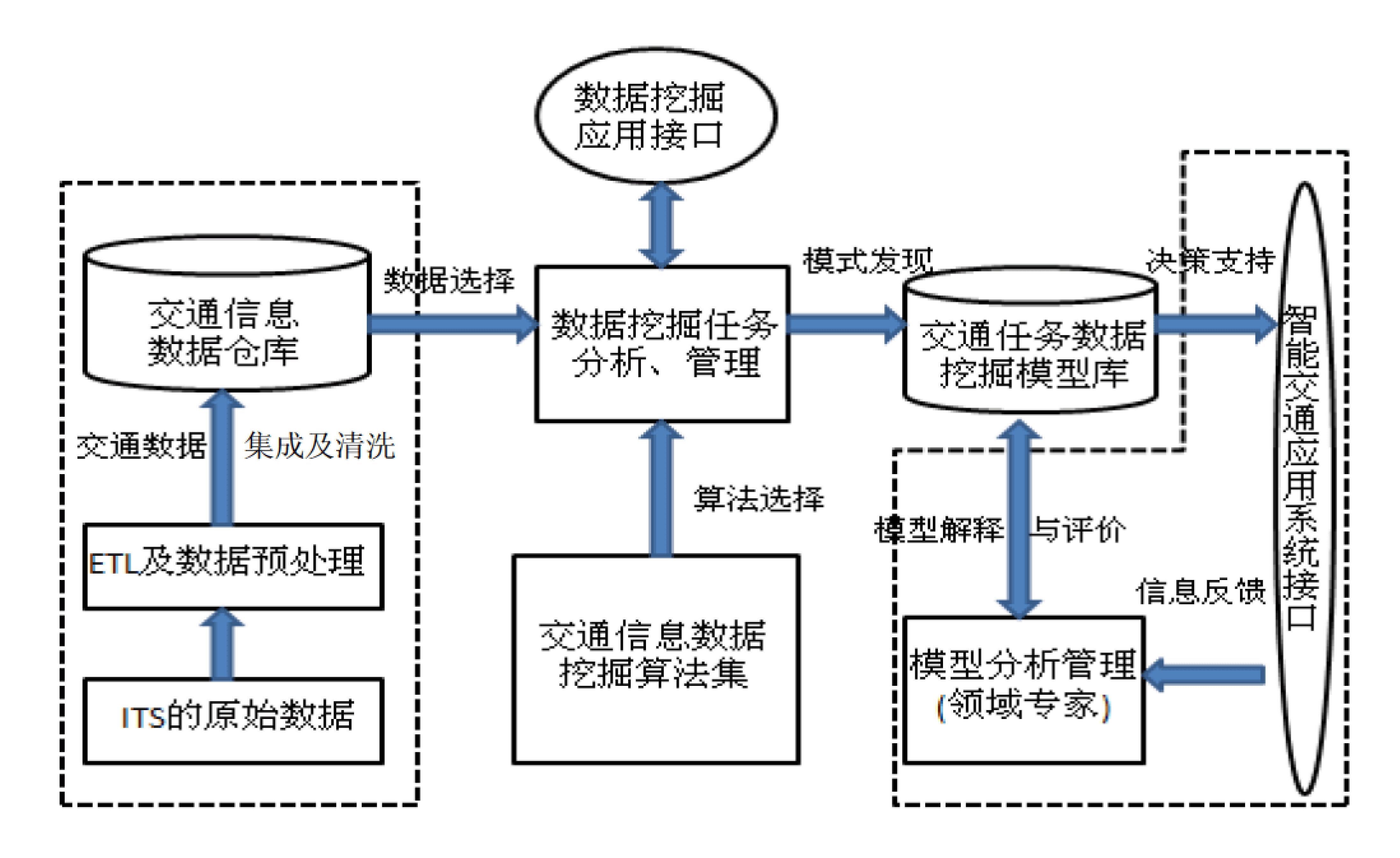
徐教授回应说：“对于这一问题，利用聚类分析方法，对分布在道路网络空间中的、环形感应线圈检测器检测到的交通流数据进行空间聚类分析，使具有相似性质且具有空间关联性的交通流数据对象聚成一类，可以发现道路交通流的空间分布模式。比如，通过基于凝聚的层次聚类算法思想，设计高效的交通流空间聚类算法，自底向上生成道路交通流的空间聚集类。”

“那如何有效地对交通流量进行预测呢？”刘队长问道。

“在空间聚类的基础上，我们利用流量序列相关性来预测交通流量，用基于神经网络方法实现道路交通流状态的预测。”徐教授说。

“能够进行有效地交通流预测后，我们的智能交通系统如何建立呢？”刘队长又问。

徐教授解释说：“智能交通数据挖掘应用平台主要划分为四层：数据层、数据挖掘算法工具层、分析逻辑层和应用系统层。其中分析逻辑部分，以交通流量预测为例，包括交通流序列相关性分析、交通流序列分割等分析模型。四层体系结构以数据挖掘算法工具为核心，在数据挖掘算法工具层和数据挖掘应用系统层之间增加分析逻辑层。在分析逻辑层抽取特定分析所需要的分析模型，并映射到合适的数据挖掘算法和分析流程。做到数据挖掘技术与具体应用紧密结合。”



“那么在此基础上如何构建智能交通系统呢？”刘队长又问道。

“由于课堂时间有限，我只给大家简单介绍下智能交通系统的基本框架结构，请大家看大屏幕！”

徐教授走上讲台，拿起激光笔指着屏幕介绍说：“智能交通系统的体系结构分为四个部分，分别是交通信息数据仓库、智能交通数据挖掘任务分析管理及算法、交通任务数据挖掘模型库及与智能交通系统的应用接口。智能交通系统的体系结构实现了交通流量的短时预测以及前期数据清洗、交通道路状态判别、交通事故数据的关联分析等交通领域内基本的数据挖掘需求。”

“哦，原来是这样。有了智能交通系统我们就可以有效预测某时刻或者某时间段的交通流量，指导调度计划，同时也可以指导交通线路的改善、改建。”刘队长感慨道。

此时，下课铃声响起，徐教授说道：“感谢刘队长的配合，今天的课程就到此结束。”

第 5 章　数据挖掘在冶金行业的应用

大家陆陆续续走进教室，发现徐教授坐在讲台上熟悉教案。学员们赶快打开笔记本，做好上课的准备，不再像以前那样寒暄一阵。

上课铃声过后，徐教授微笑着环视了一下教室，看到大家都到齐了，说：“大家好！今天我们要讨论的内容是数据挖掘在冶金行业的应用。人所共知，冶金生产属于流程工业。李部长，你就先介绍一下什么是流程工业吧……”

5.1 流程工业这点儿事

李部长在 T 钢铁公司干了近二十年了，提起他的老本行，有说不完的话。

他走上讲台："流程工业是指生产连续不间断或半连续批量生产的工业过程，如炼油、化工、电力、冶金、造纸等行业，其共同特点是工艺流程基本不变，但生产周期长，生产过程复杂，工艺参数特别多。"

"那我国流程工业的现状如何？"徐教授引导着李部长的话题。

李部长意味深长地说："前些年，我国流程工业企业普遍存在着能耗大、产品质量差、生产工艺落后、自动化及操作水平低等问题。近十年来，通过不断引进、消化吸收国外的先进生产线，研制具有独立知识产权的生产设备，情况有了极大的改观。比如我们公司年生产规模由原来的 200 万吨上升到 1050 万吨的主要原因是我们引进了清一色的德国装备。德国装备是全世界上最先进的，自动化程度相当高。我们实现了真正的'数字化'钢铁。"

"引进、消化、吸收、再创新，这是每一个国家发展的必由之路。只是我们有中国特色的速度，比别人跑得快而已。快当然是好事，但可能'消化不良'。"徐教授继续引导话题。

"确实，大部分公司与我们一样，存在着对先进设备驾驭能力不足的问题。"李部长如实说。

"李部长，你不是说你们引进的全是世界顶级的洋玩意儿，那么操作人员只需按按电钮、敲敲键盘就行了，还有什么困难的事情？"税务局姚局长调侃道。

一向冷静的李部长情绪有点激动了："此言差矣，姚局长！税务管理你内行，但隔行如隔山呐！你有所不知，流程工业有了最先进的生产设备也只具备 80%的生产能力，另外 20%就是使用设备的软实力。这 20%才是使竞争利于不败之地的关键，它比

那 80%更重要。因为 80%，你有，我有，他也有。而这 20%，却不是谁都可以花钱买来的。”

看到李部长的话落到了自己摆的“龙门阵”，徐教授顺水推舟：“好了，李部长，你还是具体给大家介绍一下 20%的软实力吧。”

李部长来劲了，大声说：“同样的先进设备，我们的能耗为什么比国外优秀企业高出 3%～8%？我们的炉温命中率为什么比别人低 2～5 个百分点？我们的板材侧翻为什么比韩国宽 2～5 毫米？我们的不锈钢成本平均每吨比日本高出 20 多美元？我们的钢材夹杂、重皮为何比欧美国家的钢铁企业严重？”

李部长擦了一下脸上的汗水，继续说道：“同志们，听到这些，你们肯定与我一样急呀！发达国家能够充分发挥‘数字钢铁’系统的作用。应用数据挖掘技术，对生产过程不断优化，使数据变成了黄灿灿的‘金子’。而我们的数据每天以 3GB 的速率增加，可这些数据却躺在昂贵的信息化系统里面睡觉！”

R 钢铁公司的何总也深有同感：“李部长说得中肯，差距就在这里，有了先进的设备，还需要结合生产的实际情况不断地优化工艺过程！”

徐教授由前排回到了讲台，说道：“流程工业的生产过程优化问题按时期可分为两大类，一类是产品设计时的优化，工艺流程、生产操作条件根据需要都可以调整，但确定之后一般不作修正。另一类是运行中的优化，此时工艺及设备因素均已确定，只有操作条件可以变动。当环境条件变量变动时，运行变量就需要及时作相应的调整。”

徐教授环视了一下教室，发现大家有点迷惑，接着解释说：“一般而言，生产装置的操作条件在设计时已按设计指标优化并定义为标准操作参数，如原料配比，过程单元设备的工作参数等。但在生产过程中，由于原料成份、参数命中率的波动、设备老化等工况条件的改变等，设计阶段定义的操作参数往往不能达到期望的效果。”

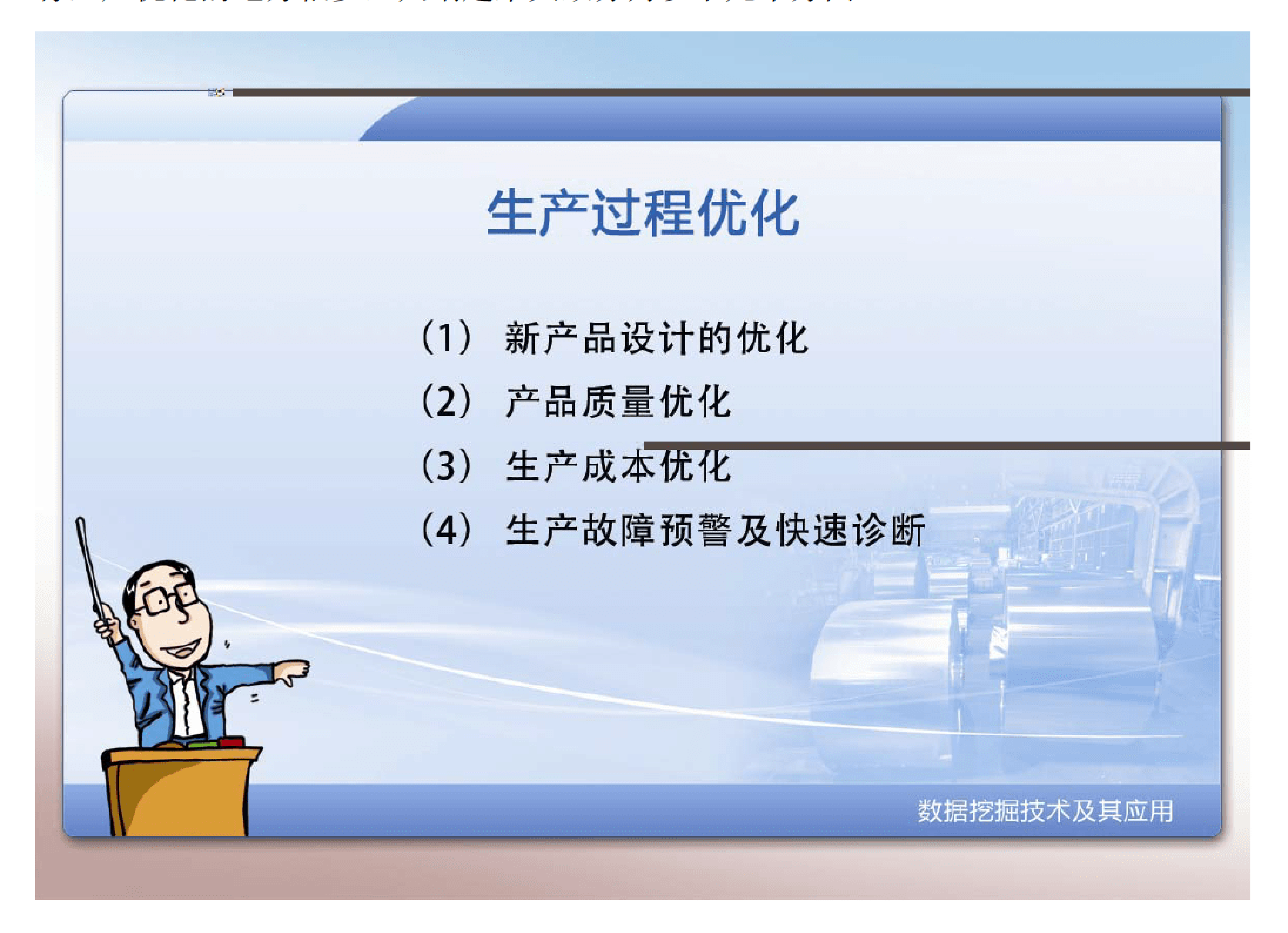
“那操作优化能带来什么好处呢？”有人问道。

徐教授解释道：“操作优化的目的是在现有工艺及设备条件下，通过调整可控变

量，使生产过程处于最优工况邻域，从而提升质量指标、提高产量、降低能耗。”

R 钢铁公司的何总问道：“那需要从哪些方面进行生产过程的优化呢？”

徐教授回答道：“流程工业是多工序的复杂生产过程，可以利用数据挖掘技术进行生产优化的地方很多，归纳起来大致分为以下几个方面。”



“徐教授，您先给我们讲讲如何进行新产品设计的优化？”电力公司的刘经理请求说。

“新产品开发是现代企业竞争的重要体现。新品试制通常需要作大量实验。如能缩短新产品的研制周期，就能为企业带来较大的经济效益。通过在实验过程中收集的数据，利用数据挖掘方法建立数学模型，能够较快地达到研制目标，使新产品更快地投产。”徐教授在讲台上一边踱步一边说道。

“徐老师，产品质量历来是企业永恒的主题，怎样进行产品质量优化呢？”R 钢铁公司的何总也问道。

徐教授扭头转向何总，亲切地回答道：“如果能在产品产出之前，通过一定方法根据生产参数估计出产品质量指标，我们就可以调整输入参数，保证生产输出指标控制在目标范围内，最终就可以少出废品。利用生产过程积累的数据，通过机器学习方法建立产品质量指标和其影响因素之间的函数关系，以及研究如何调整这些参数，从而可以提高产品质量。”

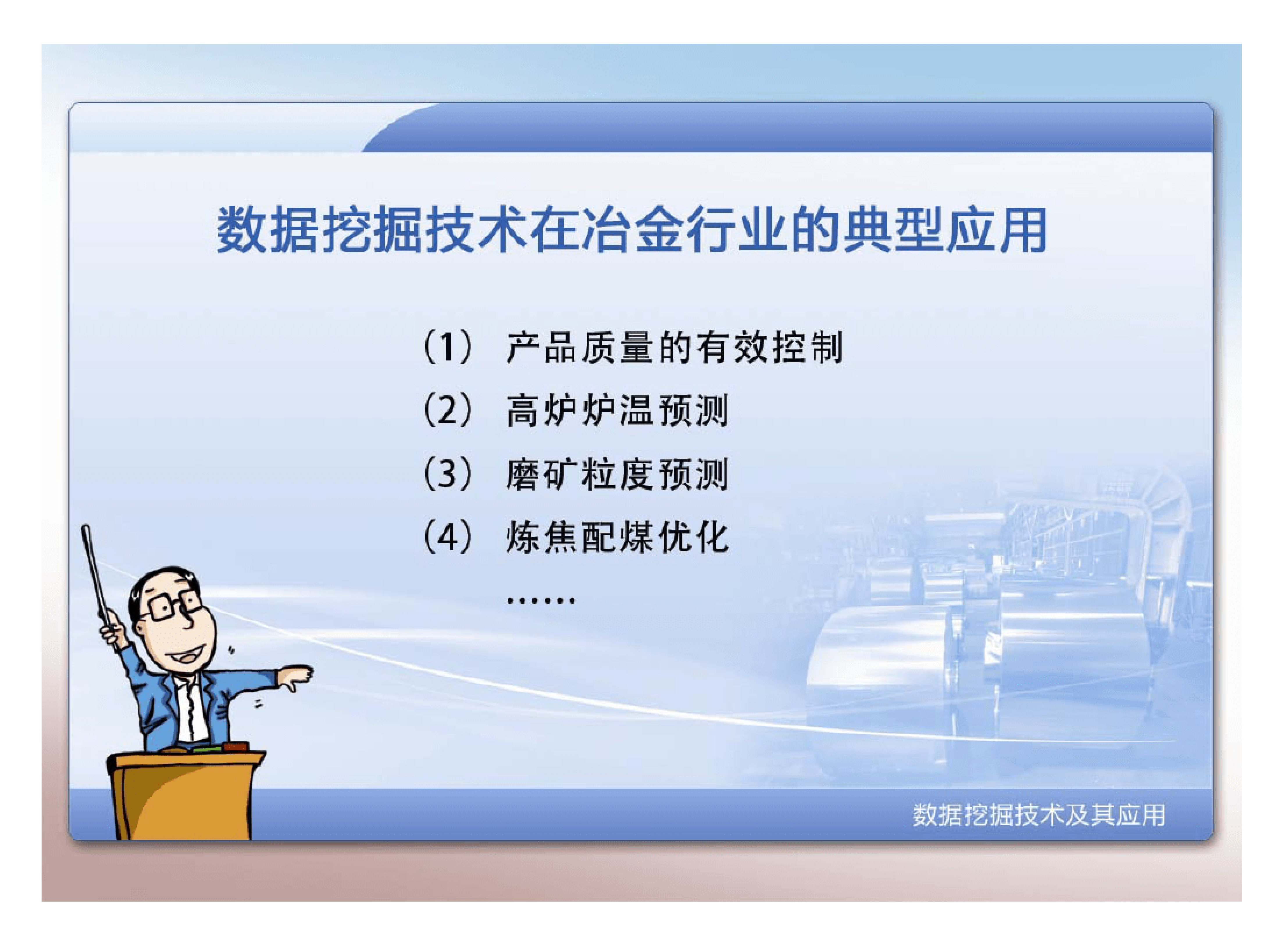
徐教授扫视了一圈台下的学员，喝了口水继续讲道：“产品质量和信誉是现代企业的生命线，许多产品的质量问题要在长期使用中才能显露出来。为了保证产品质量的可靠性，必须把好产品检验关。如何能从短期测量察觉产品的长期性能？这也需要通过数据处理，找出短期测试指标和长期使用特性的关联，建立数学模型，使产品检验更加有效。”

“生产成本是企业立于不败之地的关键。徐老师，利用数据挖掘技术，怎样优化生产成本？”李部长也提出了一个问题。

徐教授回答道：“大家知道，生产成本包括原料成本、能源成本、材料消耗成本、人力成本和其他制造费用。通过长期积累的生产数据，可以学习出单位产品与这些成本项的关系，发现低成本的生产模式，从而找出降低成本的突破口。另外，通过回归和分类方法建立的产品质量指标与工艺参数的关系模型也可以找到降低原、燃料消耗的方法。生产过程难免出现故障，利用以往数据，建立预警模型和故障诊断模型，能及时正确诊断其原因，从而快速消除故障，尽快恢复生产，也可以在一定程度上减少成本。”

“徐老师，听了您的讲解，我觉得流程工业可以应用数据挖掘技术的地方真是太多了，您给我们介绍哪些方面得到应用呢？”有人问。

“大家请看大屏幕！”徐教授说。



“这节课的目的就是让大家对流程工业的数据挖掘应用有个总体的认识，下面几节课我们将探讨几个具体的数据挖掘应用问题。OK，今天的课到此结束。”

5.2 产品质量控制

徐教授走上讲台，直奔主题：“这节课，我们一起探讨数据挖掘技术在产品质量控制中的应用。”

他打开笔记本电脑，继续讲道：“激烈的国际市场竞争不断地向产品质量、新产品设计、产品成本和交货期等方面提出新的挑战。如何提高产品质量，使企业具备自己的竞争优势，已经成为企业的新挑战。随着流程工业自动化、数字化水平的不断提高，数据越来越丰富，这就为应用数据挖掘技术进行产品质量控制提供了良好契机。”

“这下我们数字化钢铁系统中的数据可有用武之地了！”R 钢铁公司的何总激动地说。

李部长的思维再次超前了一步，他说：“徐老师，记得您曾经说过，流程工业的产品质量控制问题大都可归结为机器学习范畴的分类问题和回归问题, 对吧？”

徐教授说道：“是的，流程工业生产是多工序生产，各个工序都有影响产品质量的因素。如影响钢材表面质量的因素有：元素成份含量、铸坯的厚度及宽度、控制温度、铸坯拉速、时间等。一般把这些影响因素称为输入变量，衡量产品质量的指标称为输出变量。输出变量可分为两类：一类是离散型输出变量，如板材表面是否有夹杂、重皮，纵条纹的等级等，可用 0、1、2 等整数值表示；另一类是连续性输出变量，如钢材的抗拉强度、延伸率、不锈钢边缘的侧翻等。根据输出变量的类型，前者可归结为分类问题，后者是回归问题。”

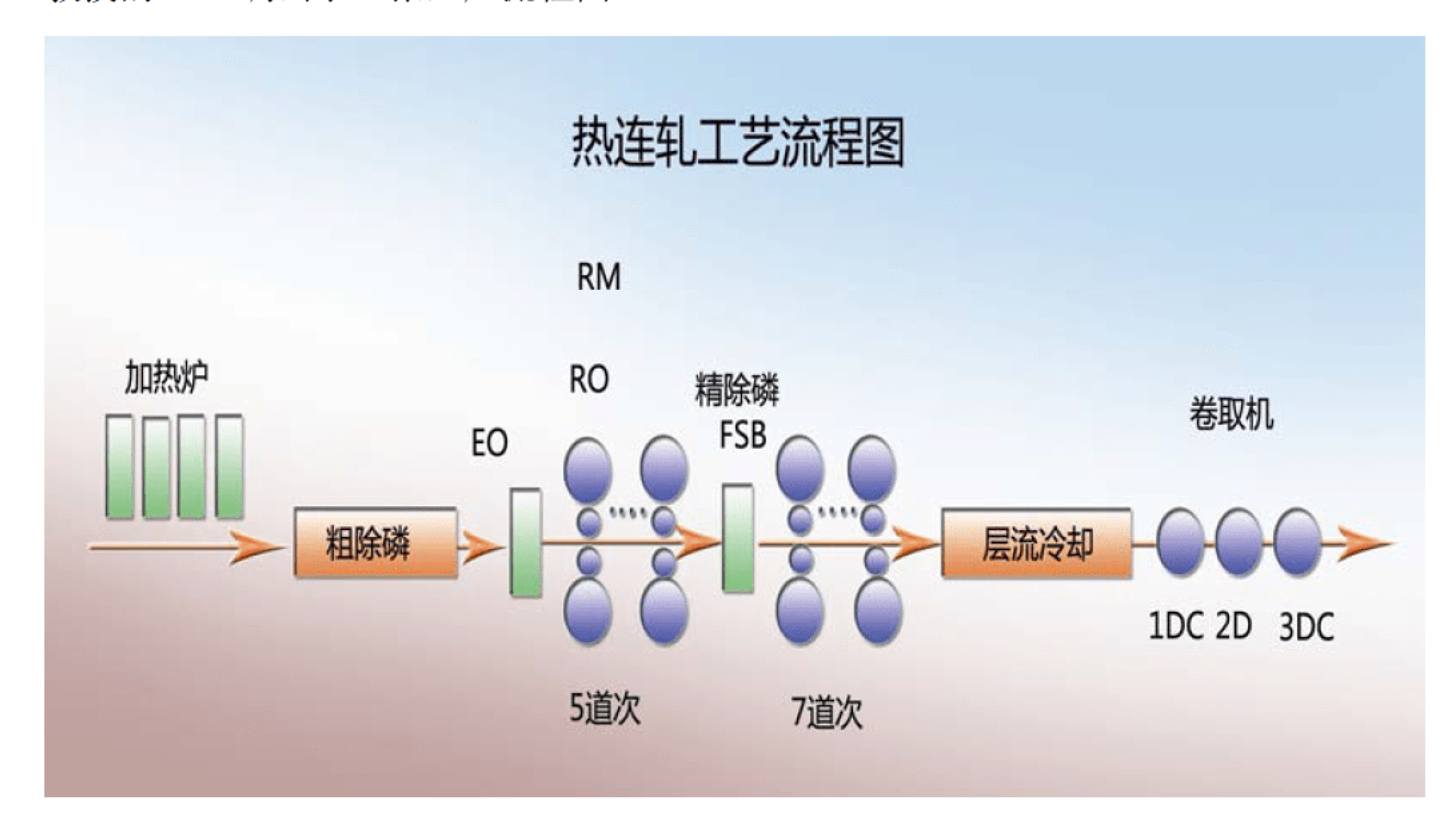
听到这里，李部长回忆起徐教授在他们公司做讲座时对产品质量控制问题的一段概括，脱口而出：“流程工业的产品质量控制问题可描述为：假定生产过程 P 的产品质量指标 y 有 p 个影响因素 y=（x1，x2，…，xp），根据对产品质量的影响因素和产品质量指标的测量数据，推断其函数关系 y=f（x，β），这里 β 为待定参数。然后根据所得到的函数，对新的工况参数，推断其对应的质量指标，这就是产品质量预测。反之，根据指定产品质量目标值反推相应的影响因素参数值，这种情况，称为逆质量问题。对产品质量预测问题和产品逆质量问题建立的模型分别称为产品质量模型和产品质量控制模型。”

“李部长的记性真好。没记错的话，这是 3 年前我讲的内容。”徐教授赞扬道。

这时，S 钢铁公司的赵总提了个建议：“这些都很抽象，徐老师以一个具体的质量控制实例给我们讲讲吧。”

“好的，就以 T 钢铁公司 1549mm 热连轧生产线板材抗拉强度和延伸率质量控制问题为例吧。”其实徐教授早有准备，答应道。

“李部长，你对这个问题再熟悉不过了，就先给大家介绍介绍情况。”说着，徐教授的 PPT 调出了一张生产流程图。



李部长从徐教授手上接过光笔，指向大屏幕说道：“企业的同志可能有同感，市场竞争是极其残酷的。我们的板材成本老是比国外同行‘略高一筹’，致使我们在国际竞标中屡屡失败。为什么呢？董事长多次召集大家研究对策，分析认为我们的设备并不比人家差，原料来源与竞争对手也没有什么区别。”

说到这里，李部长一脸无奈的样子。

“知己知彼，百战不殆。有一天，陈董事长让我把竞争对手的产品一一进行物理性能测试。结果令人大跌眼镜，国外产品的两个主要性能指标抗拉强度和延伸率竟然比我们的应标产品低了不少。”李部长继续道。

“手下败将，不服输，还自吹自擂！”S 钢铁公司的赵总与 R 钢铁公司的何总互相“咬耳朵”。

“当我把测试结果呈现给陈董事长时，他眼睛一亮，桌子一拍：‘原来如此！’”李部长像讲故事一样表情投入。

“我先是一愣，瞬间也就明白了。”李部长脸上露出了笑容。

“明白什么了？”税务局姚局长等不知所云。

“这个问题我不告诉你。”李部长风趣地引用了一句广告语。

其实 S 钢铁公司的赵总和 R 钢铁公司的何总等生产企业的学员也都明白了。

“陈董事长让我立即召开技术研讨会，拿出对策。会上，我分析道：‘在满足用户需求的情况下，降低抗拉强度和延伸率必然会降低产品成本，当然会赢得降价空间。’”李部长道出了其中的奥妙。

“现在，我们的目标明确了。首先，研究出抗拉强度和延伸率与诸影响因素的数量关系，从而按用户要求的质量指标，在尽量降低生产成本的情况下确定最优的工艺参数。”李部长进一步说。

这时，徐教授走上讲台，接着李部长的话茬讲到：“热连轧生产工艺流程包括加热炉、粗除鳞、5 道次初轧、精除鳞、7 道次精轧、层流冷却和卷取机成卷几个阶段，其工艺流程如图所示。”

徐教授在屏幕上晃动着手中的光笔，继续说道：“初步分析认为，PPT 中所示的 22 个生产变量对板材的抗拉强度和延伸率有直接影响。首先我们需要利用生产数据建立产品质量预测模型，即抗拉强度和延伸率与 22 个生产变量的函数关系，然后，应用所建立的函数关系，由用后对质量指标的具体要求反推最优的生产变量控制参数。”



S 钢铁公司的赵总问道：“徐老师，在这个问题中，建立产品质量预测模型和逆产品质量预测模型分别用什么数学方法？”

徐教授回答道：“前者采用 LASSO，即 Least Absolute Selection and Shrinkage Operator 模型，后者使用遗传算法。”

“徐老师，采用 LASSO 模型有什么优势？”R 钢铁公司的何总问道。

徐教授回答说：“前面我们说过，我们初步分析认为有 22 个生产变量对板材的抗拉强度和延伸率有影响，这些变量可能有冗余，或者可能作用很小。LASSO 模型可以在回归误差尽可能小的情况下剔除冗余变量。”

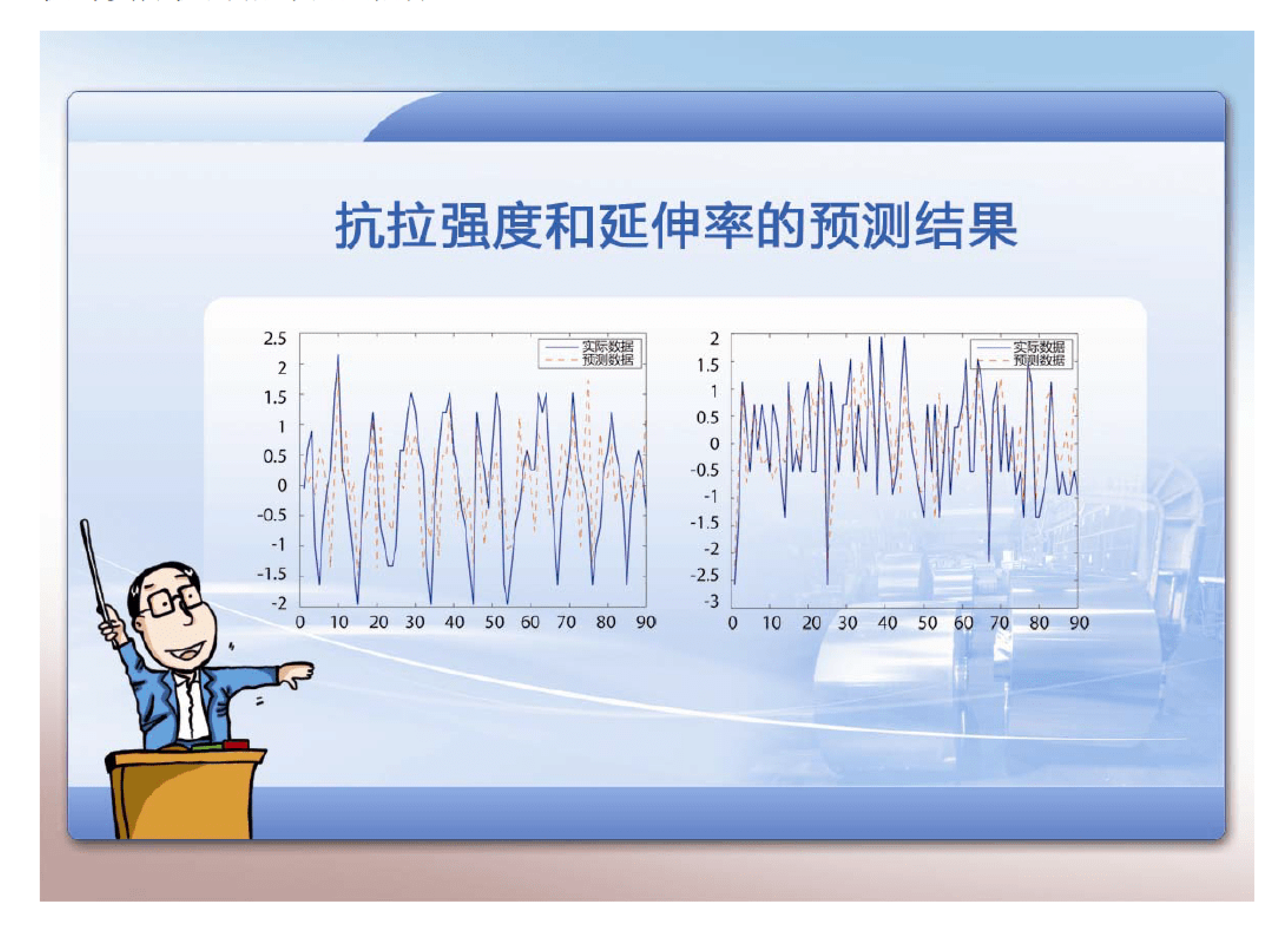
李部长经过这几年的数据挖掘实践，对模型和算法已经比较了解了，卖弄道：“常用的最小二乘、支撑向量机和神经网络等可没有这样的特点。”

S 钢铁公司的赵总问道："模型建立之后，用什么方法求解呢？"

徐教授讲解道："LASSO 模型虽然是凸优化问题，但由于使用的 1-范数是非光滑的，快速求解 1-范数正则化模型一直是人们非常关注的问题，人们提出了不少算法，对比分析后我们认为选用梯度 Boosting 算法求解。"

听完徐教授的话，台下学员嘀咕道："哦，这样得到回归模型后，任意给出一组输入参数，都可以预测出其对应的输出值了。"

徐教授接着说道："由于我们采用了线性回归模型，所以回归结果的可解释性很强。LASSO 质量模型的预测精度高，与其他方法相比，训练时间非常短，而且预测值与实际值的相对误差很低。"



R 钢铁公司的何总问道："建立了抗拉强度和延伸率的回归模型后，用什么方法由质量指标反推生产工艺参数呢？"

徐教授说："遗传算法。"

"遗传算法，听我们单位的研究生小高说是利用达尔文进化论的思想构造的一种求解复杂优化问题的算法。"一个学员说。

"是的，物竞天择，适者生存，这是自然界法则。"徐教授解释说。

"徐老师，遗传算法挺有意思，您就给我们详细讲解一下吧？"前排的一位学员请求道。

"课时太紧，我就粗略地介绍一下其基本过程吧。生物种群必须经受优胜劣汰的选择、生物进化需要染色体的交叉和变异的改良，一代一代繁衍不息，留下来的都是精品。我们就是把这个生物进化过程写成算法，就是遗传算法。"徐教授解释说。

台下一个学员激动地说："哦，我明白了！给定抗拉强度和延伸率的目标值，给出一系列的生产参数，根据回归的抗拉强度和延伸率模型，计算出这些生产参数对应的回归的抗拉强度和延伸率。留下误差小的几组生产参数，去掉误差大的生产参数。然后通过生产参数向量的交叉和变异形成一些新的生产参数，构成一定规模的新种群，然后再重复上面的过程，直到种群内有一组生产参数所对应的抗拉强度和延伸率与期望的目标值足够接近或进化出了一定的代数为止。"

可能是有点不好意思，S 钢铁公司的赵总小心翼翼地问道："徐教授，热连轧产品质量控制的过程我都听明白了，那怎样评价模型的好坏？"

徐教授解释道："通常以命中率来衡量逆质量控制模型的优劣。如果某一组输入输入对应的输出值与真实值的相对误差小于 5%，模型在这一组输入上命中，则此样本点称为模型的一个命中点。命中点总数占总样本数目的百分比称为模型的命中率。"

S 钢铁公司的赵总又问道："模型的命中率如何？"

徐教授说："在逆向质量控制时，应用遗传算法建立质量控制模型，回避了逆向质量模型的存在性和唯一性问题，计算的结果达到了精度要求。"



5.3　高炉炉温预测

徐教授看到大家都已做好上课的准备，开门见山地说："今天我们来一起探讨数据挖掘技术在高炉炉温控制中的应用。"

"赵总，听李部长说，你是 S 钢铁公司的炼铁专家，你先给大家介绍一下高炉炉温控制的作用吧。"

赵总向讲台走去，毕恭毕敬地说道："前天徐教授就给我布置了任务，让我介绍高炉炉温预测的基本知识。我准备了一下，希望对大家有所帮助。"

"高炉是横断面为圆形的炼铁竖炉，其主要作用是用化学和物理方法减少铁的氧化物含量，产出优质铁水。随着计算机技术的发展，大规模非线性数据处理成为可能，高炉炼铁从最初追求规模效应，逐渐走向强调高炉的长期稳定、顺行、高产、低耗。在高炉炼铁过程中，炉温是高炉控制最为重要的一个指标，准确控制炉温并维持炉温的稳定对高炉炼铁生产具有特别重要的意义。"赵总概括了高炉炉温控制的目标。

赵总满脸是汗，补充了点水分，继续说道："由于生产是连续进行，炉内温度很高，很难直接测量得到。而硅元素的还原速率受炉内温度与热量影响的灵敏度远比铁高，因此通常就用铁水中的含硅量来代表炉温。硅含量越高，炉温愈高。"

"通过预测高炉内硅含量来预测炉温，妙，实在是妙！"一个学员饶有兴趣地感慨道。

"高炉炉温预测有其特殊性，赵总，你给大家再介绍一下高炉冶炼包含哪些控制参数，以便我们选择合适的数据挖掘技术。"徐教授说道。

这个可难不倒名副其实的炼铁专家，他滔滔不绝："高炉炼铁包括配料、上料、布料、鼓风、富氧喷煤、出渣、出铁等过程。它们之间互相作用和影响，各个环节的影响参数多达数百项。高炉体内流体存在复杂的相态，煤气、炉料、渣液三项之间不断进行着动量、质量和能量的传递和转换。"

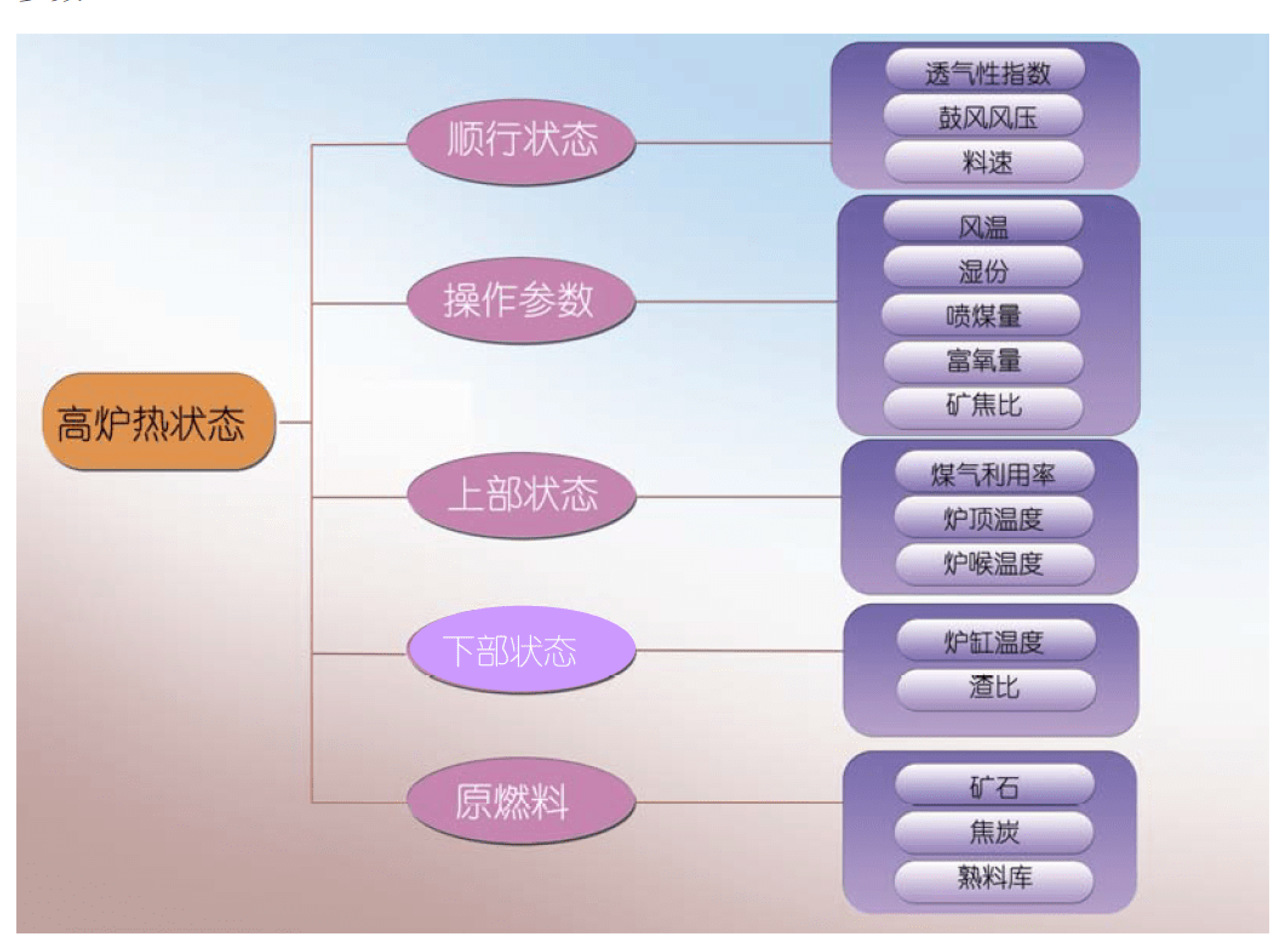
"小小的高炉，内部真是翻江倒海般热闹，看来对其进行数据建模那是相当的困难。"台下有人说道。

赵总意味深长地点头，回应说："是的，高炉炼铁过程极其复杂。在冶炼过程中，固体下落同时气体要上升，如果气体对固体的阻力太大，就会导致悬料等炉况事故，而若炉内化学反应状态不均匀，又将导致崩料等炉况事故。总之，高炉冶炼过程所具有的时变、高维、多频、分布参数等复杂特性和封闭条件下的操作，都使得参数检测

非常困难。最终导致高炉炉温的建模和控制变得非常困难，成为冶金自动化领域的技术难题。”

“难题并不可怕，可怕的是不能与时俱进，抱着老方法不放，不能充分发挥机器学习研究的新成果。”徐教授趁机向在座的公司领导敲警示钟。

赵总打开了他的 PPT，继续说道：“高炉生产过程是一个非常复杂、高度耦合的非线性过程。高炉内影响铁水硅含量的因素很多，大体上分为两大类：状态参数和控制参数。状态参数是指反映高炉冶炼过程状态的参数，同时也是控制参数作用的结果，它们无法像控制参数那样进行实时调整。大家请看影响铁水硅含量的控制参数和状态参数。”



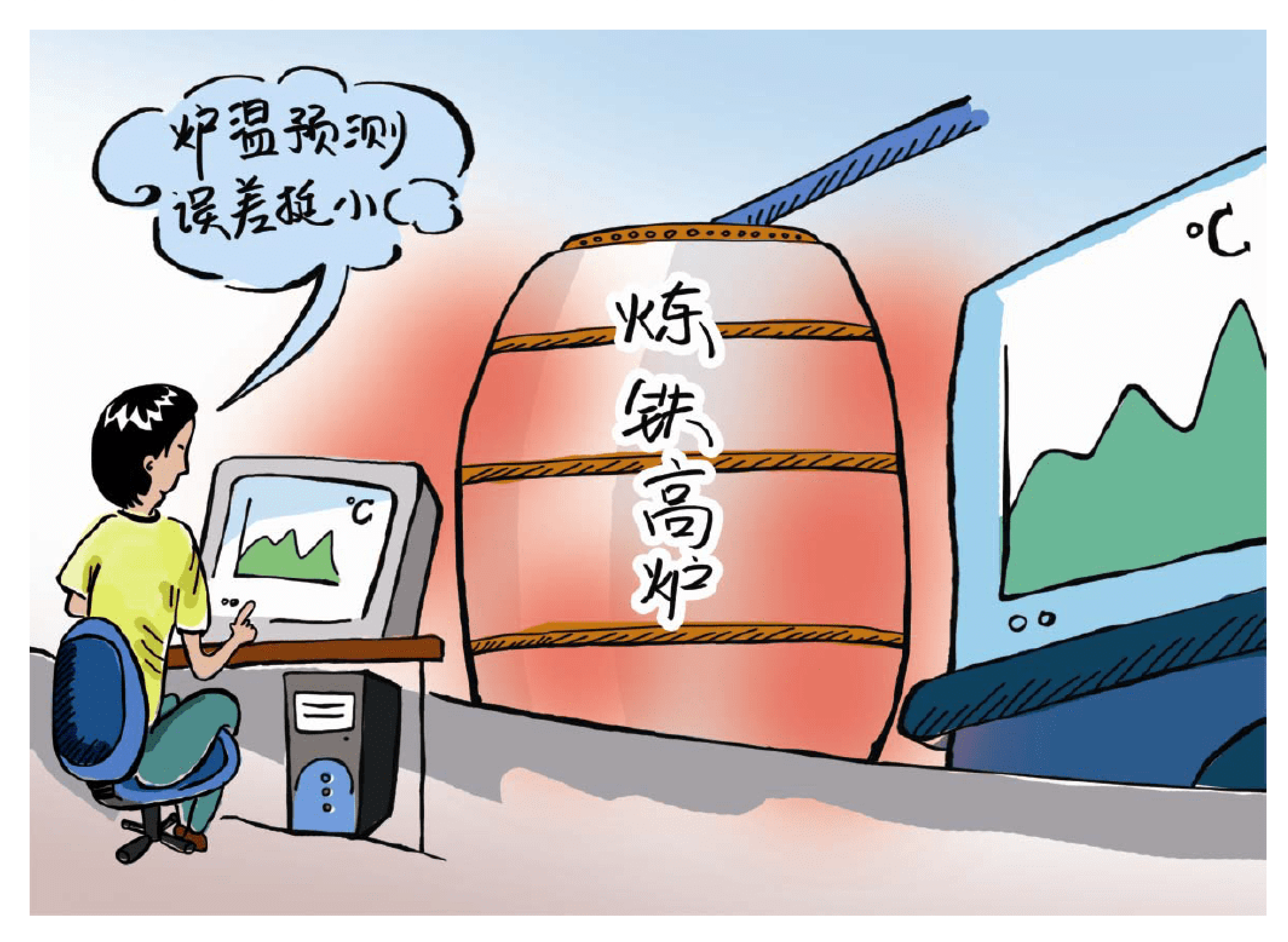
台下半晌无言，良久，有人低声道：“这么多参数，都看晕了。”

徐教授示意赵总休息休息，走上讲台回应道："是的，影响高炉铁水硅含量的高炉状态参数和控制参数很多。在建立铁水硅含量预报模型时，将其全部都作为模型的输入变量势必会增加模型的复杂度，且会影响实时速度。"

"那么我们如何来选择较为重要的状态参数和控制参数呢？"R 钢铁公司的何总问道。

"一方面利用高炉操作人员的经验进行选取，另一方面通过这些参数与铁水硅含量的相关性进行分析，选择相关系数较大的参数。"徐教授简明扼要地回答道。

R 钢铁公司的何总道出了硅含量预报的最大难点："由于高炉炼铁过程所具有的慢时变特性，根据历史数据得到的铁水硅含量预报模型在现场运行一段时间后，模型往往会失效，有什么好的方法解决这一问题？"



徐教授答道："我们采用增量式支撑向量机技术进行高炉铁水硅含量的在线建模。"

何总思前想后，自觉无法明白个中玄机，只好请教道："前面我们所学习的支撑向量机与增量式支撑向量机有什么差别？"

徐教授和颜悦色，立即解释说："增量学习是指在原来的学习样本情况下，增加新样本的再学习方法。这种学习方法有着明显的优势，一方面由于在新训练过程中，充分利用历史的训练结果，从而显著减少了后继训练时间，同时对于高炉炼铁过程这类渐变问题，新样本所提供的信息与历史数据所提供的信息量是不同的；另一方面增量学习过程将舍弃无用的样本，无需保存全部历史数据，减少对存储空间的占用，可以运用于在线学习中。"

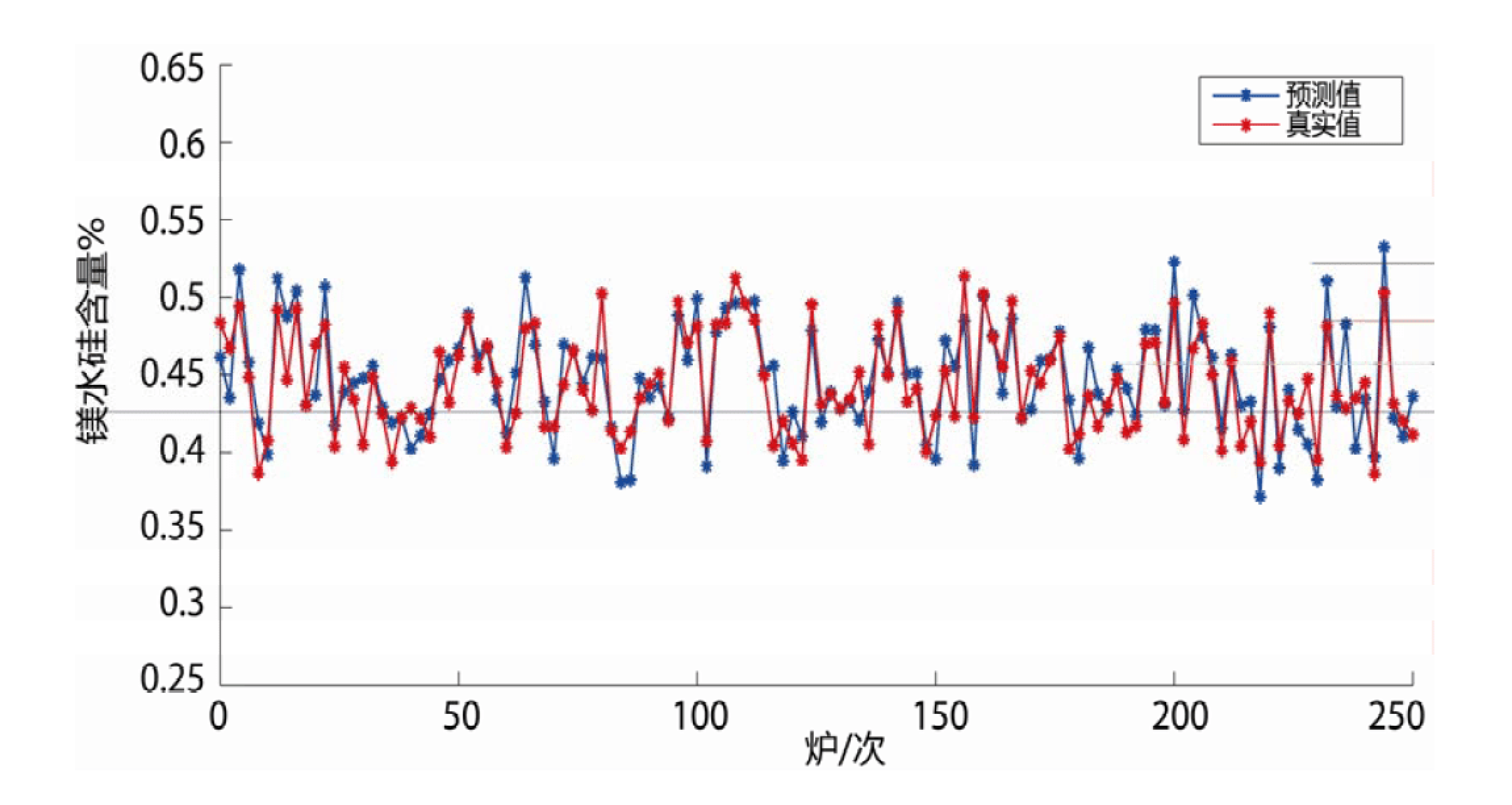
何总穷追不舍，笑眯眯地道："增量式支撑向量机学习具体怎么训练模型呢？"

徐教授扶了扶眼镜，说道："SVM 增量学习基于以下两点进行：由于在整个训练样本集得到的训练回归函数与只在支撑向量上训练的结果是一样的，我们就可以用相对较少的支撑向量代表整个训练集；其次增量样本中如果属于非支撑向量集合 R，即在回归间隔线之内的点，将其加入工作集中，不会改变训练结果，当增量样本不在间隔线之间时，则将改变支撑向量机的回归函数。"

"原来如此，明白了！"R 钢铁公司的何总摸了摸脑袋，恍然大悟道。

而大部分学员感到疑惑不解，茫然地摇摇头。

接着，徐教授解释道："现在高炉铁水硅含量预报模型大部分是离线建模，我们采用基于增量型支撑向量机高炉铁水硅含量的在线建模。增量型训练过程中，可以在高炉铁水硅含量预报模型中不断增加能够代表新工况信息的样本，同时控制工作样本集的规模。它真正实现了在线预测，而且预测精度远远高于离线模型。请看增量式支撑向量机高炉铁水硅含量的在线建模效果。"



5.4　磨矿粒度预测

一上课，徐教授便问道："我国采矿最早可以追溯到石器时代对石器材料的选取。后来，随着冶金业的兴起，采矿和选矿技术也逐渐发展起来。在采矿的过程中，需要使用哪些技术？"

冶金业内的贾总抓了抓头皮回答道："这个得看矿体埋的深浅程度，那种比较浅的'草皮矿'或'鸡窝矿'，只要通过露天开采技术，把表土或薄层岩层剥除，掘下数尺就可得矿。那种比较深的，需要用到立井开采技术。"

犹豫了几秒钟，台下另外一个学员补充道："碰见岩石类的，会使用岩石破碎方法，主要有工具破碎和火爆破碎两种方法。开凿坑道时，可用工具破碎岩石，或者火加热岩石，使岩石内部结构受到破坏。"

铁路局的高局长忽然意识到了什么，也说道："在明代，矿井深度就达到数百丈了。像这种考虑井巷通风、排水，矿石原料提运，的复杂工程从安全角度考虑，必须

建设矿井下面的支撑保护，根据土壤和矿石的特点，来判断矿井地压方向，采用留石柱、木架、充填支护。”

徐教授眨了眨眼，接过话茬：“说得都很好，我们都知道采矿之后的环节就是选矿，选矿也是冶金前一个非常重要的环节。说起选矿，就不能不说磨矿，大家对磨矿的了解有多少呢？”

贾总有些害羞地说道：“磨矿是选矿生产过程中前期不可或缺的一个环节，其目的是将大颗粒矿石破碎到一定程度。”

另外一个学员咂了咂嘴巴，补充道：“磨矿将有用矿物和脉石矿物分离，呈单体解离状态，以利于有用矿物的选别。”

看到大家积极发言，鼓风动力集团的王总也赶忙响应道：“磨矿流程是选矿厂投资最多、电耗与钢耗最高的生产工序，其生产过程的优化控制可稳定产品质量、提高磨矿效率、降低能耗，并且直接制约着选矿产品质量和金属回收率。”

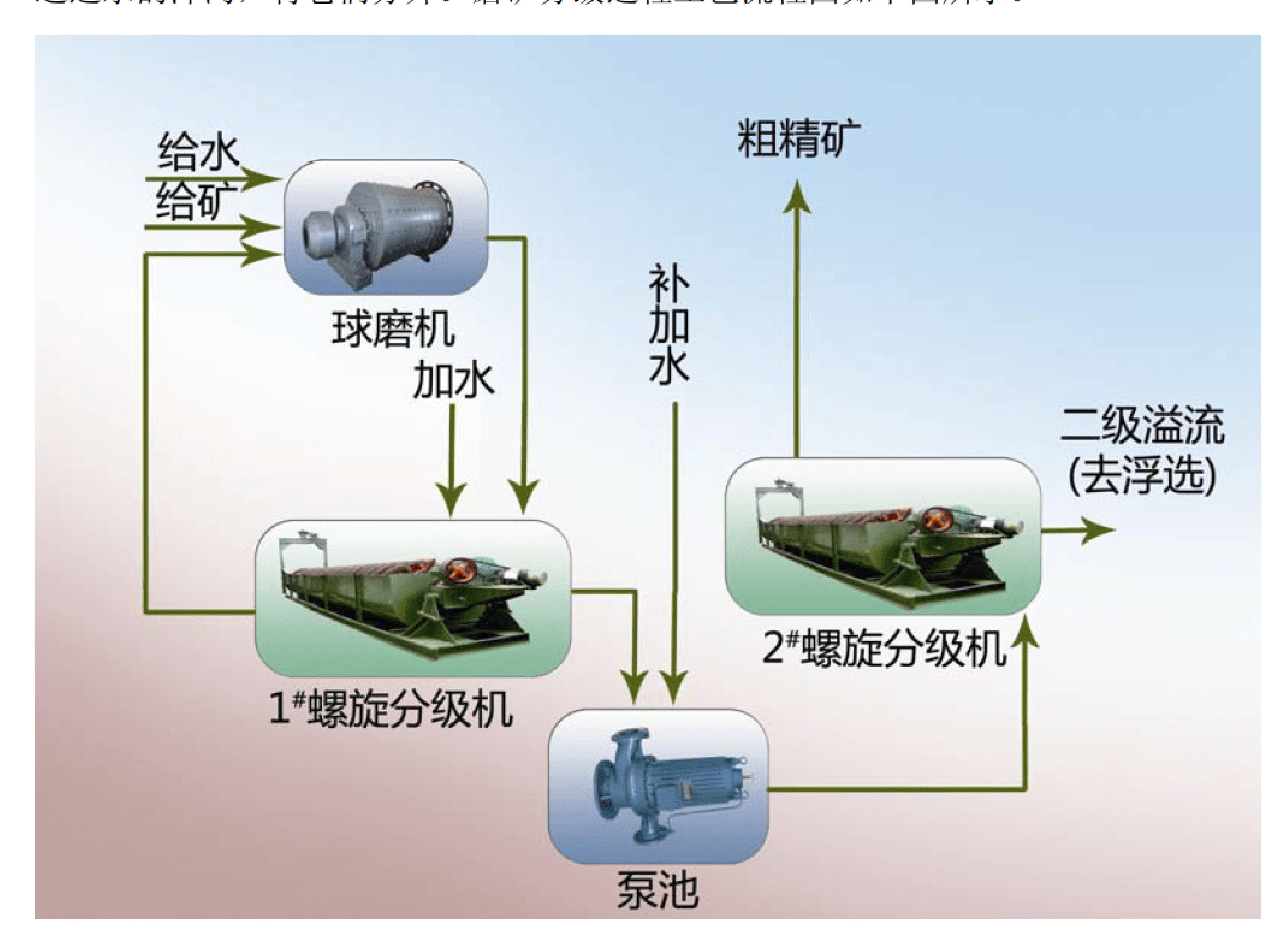
台下有人开玩笑说道：“磨矿？不太知道。不过我对这个感兴趣，回头可以自己磨矿冶炼金子。”

徐教授继续说道：“我国明朝的时候就有磨矿的记载了：先将破碎矿石，再用碓舂成细末，然后用大桶盛水，把矿末投入水中搅拌，搅后，浮在水面上的称细粘，悬浮水中的称梅砂，沉于桶底的称粗矿肉。再将细粘和梅砂、粗矿肉用尖底淘盆或者舟形木盘淘洗，取得精矿。”

电信的冯总洋洋自得地说道：“那个淘床一般是木料的，四周有边，淘床上固定一个圆竹筐。将沙倒入筐内，手把住淘床后面的木架，不住掀簸，用水频洗沙筐，则沙随水流，金从筐底细缝透下，沉淀于淘床上。淘床两头镶板，中空三尺多，另安木板一块，上面横刻木槽，筐底透出的沙金顺水沉入槽内。另用木匣一个，空出一面，类似簸箕形状，将槽内矿质扫入木匣，在水中淘洗。”

“徐教授、电信的冯总，你们俩讲得太生动了。那个磨矿的画面就在眼前了”，玻璃公司的尚主任歪着脑袋说。

徐教授笑着，继续说道：“随着现代技术的提升，采矿和选矿技术也发展迅速。对比古代磨矿技术，其原理没有发生本质变化：利用岩石砂粒与矿物颗粒的比重不同，通过水的冲淘，将它们分开。磨矿分级过程工艺流程图如下图所示。”



贾总自我推荐，主动给大家讲道：“看图说话第一部分：给矿机和一定比例的水，同时将原矿给入 I 段球磨机进行研磨，研磨后的矿浆排入分级机，同时在分级机的入口加水。分级机返砂送入 I 球磨机，分级机与一段球磨机形成回路。分级机的溢流进入泵池，同时补加水进入泵池。”

工行的张行长有样学样，也凑热闹地说道：“看图说话第二部分：胶泵将泵池内的矿浆以一定的压力和浓度打入水力旋流器，矿浆在旋流器内部得到分级。粗粒级矿浆由旋流器底部的沉砂口排出，形成循环负荷，进入 II 段球磨机再磨。磨后的矿

浆经磨机排放进入泵池，进行下次分级。细粒级矿浆由旋流器顶部溢流口排放，形成二段球磨溢流矿浆进入选别工序。”

听完详细的流程后，大家都有了更深刻的理解和认识：“那确实，和过去的机理没发生什么本质的变化。”

看大家都理解到位了，徐教授进一步说道：“磨矿中，矿质的粒度不但是磨矿作业最重要的生产指标，也是影响后续选别作业的精矿品位和回收率的关键因素。”

看大家茫然的表情后，徐教授示意贾总进一步给大家解释。贾总便站起来说道：“磨矿粒度过粗时，矿石中有用矿物与脉石之间没有充分解离，存在大量的连生体，难以把有用矿物选出来；反之，会使有用矿物产生泥化，同样不利于选别并且增加能源消耗。因此，保证合适的粒度是磨矿过程控制的关键。”

徐教授喝了口水后说道：“由于在线粒度分析仪过于昂贵、维护保养复杂，选矿厂难以承受，且现有的粒度计检测周期长，不能满足实时控制的要求。这些情况都使磨矿细度方案的确定有很大难度，传统的统计分析方法在磨矿细度问题上也显得力不从心。”

贾总接着说道：“生产中矿石以及辅料的种类越多，越难以把握最优磨矿细度。矿石中矿物种类之间的相互作用和交叉影响，使磨矿细度具有很强的非线性特征，难以进行单因素分析。”

接过贾总的话题，台下有人说道：“在实际生产中表明，载金矿物与脉石矿物的共生关系较为密切，并且矿物种类较多，如：褐铁矿、黄铁矿、白云石、石英、石墨、白铅矿等，所以磨矿细度要达到一定的程度才能使有用矿物单体解离。”

贾总说道：“是啊，多次的磨矿细度试验发现，在磨矿细度增加的初期，矿石品位有所下降，但回收率有缓慢上升的趋势，当细度达到一定程度时回收率不再上升，矿品位也处在较高位。”

“就是，数据挖掘能为磨矿献什么样的计、什么样的策，徐教授，我已经迫不及待想知道了！”台下另外一个学员听了前面的铺垫介绍后，心急地问道。

徐教授说道："听我慢慢讲，太快了大家会囫囵吞枣，这不是个好现象。让我们回头看磨矿流程工艺图，可以发现该过程中旋流器溢流矿浆粒度为整个磨矿作业的关键工艺指标，关系到磨矿作业的质量，直接影响后续选矿过程的金属回收率和精矿品位。因此，对其溢流粒度的有效控制显得至关重要。"

"徐教授，我们要怎么预测磨矿粒度呢？"台下有人发出疑问。

徐教授解释道："首先找到与磨矿粒度相关的变量，如球磨机给矿量、给水量、磨机电流、I 段螺旋分级机溢流浓度、泵池补加水量、II 段螺旋分级机给矿浓度等数据。"

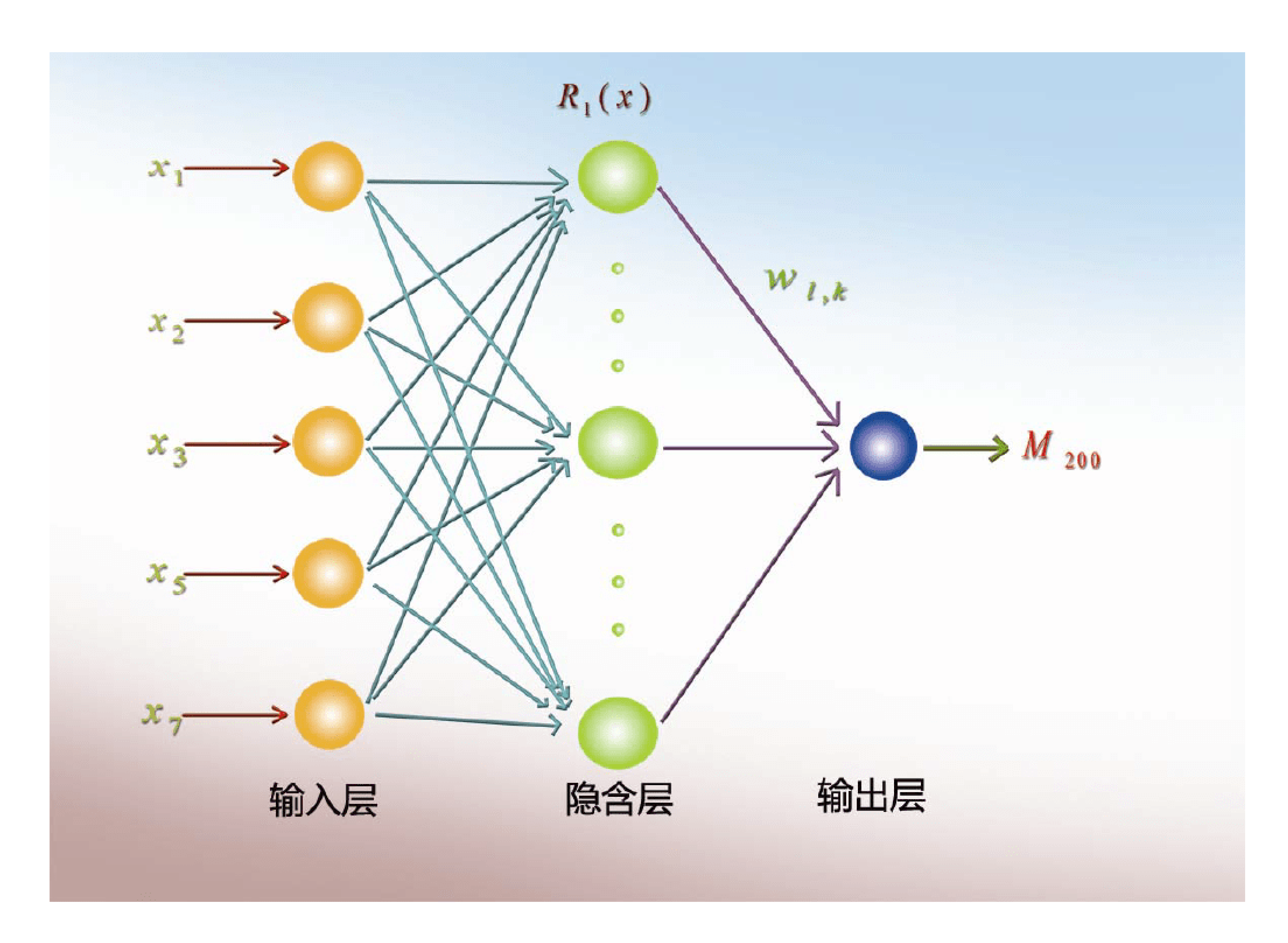
变量名称	变量类型	是否选入建模
X_1：给矿量	自变量	是
X_2：给水量	自变量	是
X_3：磨机电流量	自变量	是
X_4：返沙比	自变量	否
X_5：I 螺旋分级机溢流浓度	自变量	是
X_6：泵池里补加水量	自变量	否
X_7：II 螺旋分级机给矿浓度	自变量	是
M_{200}：磨矿粒度	目标变量	是

"徐教授，表中的最后一栏中的建模选取或不选取标准是什么呢？"台下有人问道。

徐教授回答道："首先通过对收集磨矿细度数据，对其进行滤波、去量纲等预处理操作。其次，采用数据挖掘技术中聚类、主成分分析等方法，就可以找出在此阶段中影响回收率的关键因素。以影响回收率和冶金品位的关键因素作为自变量，回收率为因变量的神经网络预测模型。"

玻璃公司的尚主任问道："神经网络进行预测，比回归的优势在哪里呢？"

徐教授解释道："由于神经网络具有很强的并行处理、自适应、自组织、联想记忆及容错能力，所以可以为复杂生产过程中难以测量的工艺参数进行在线检测提供了有效途径。同时，神经网络非线性处理能力和逼近能力强，学习时间短，网络运算速度快。"



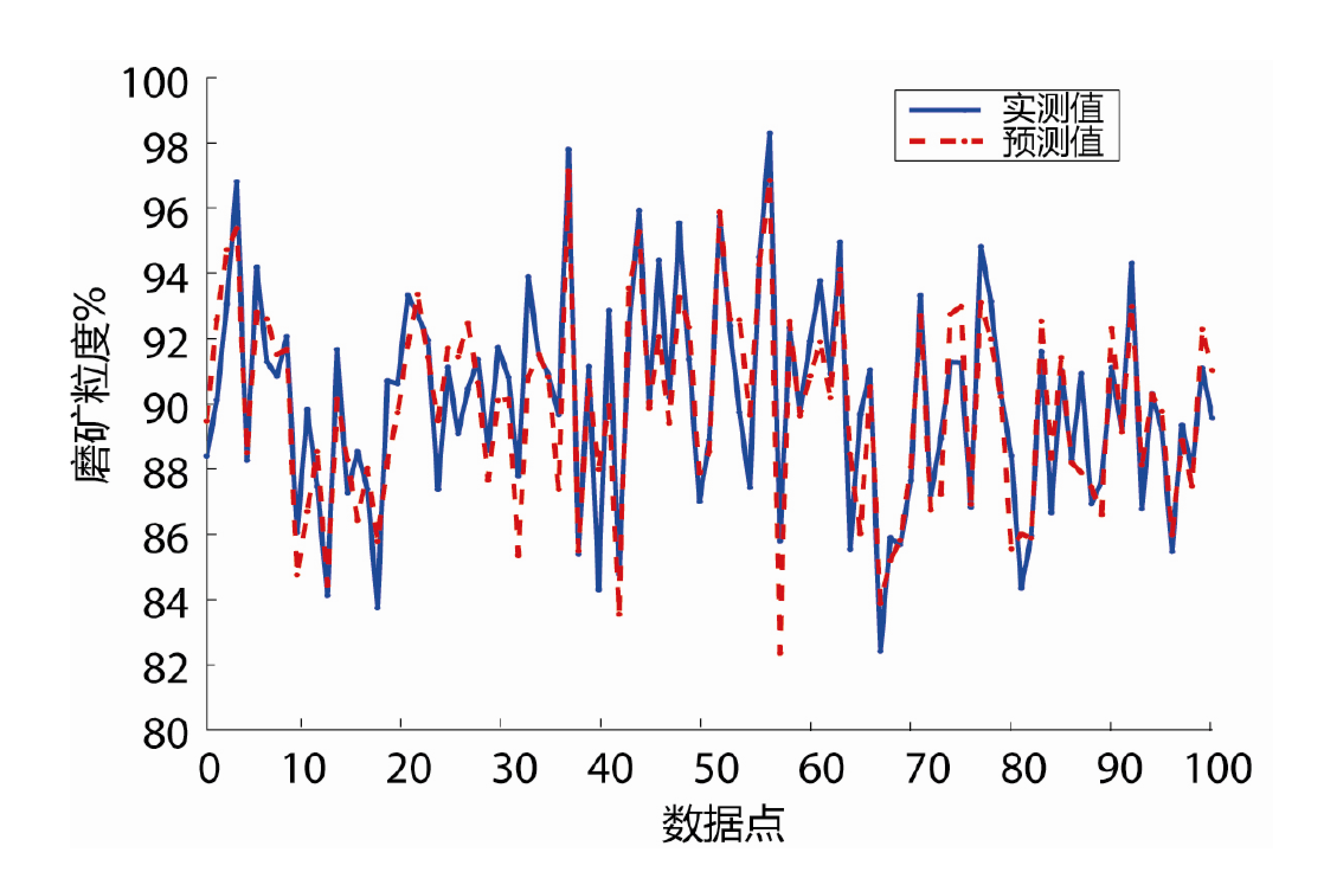
看到大家沉默，徐教授继续解释道：“采用神经网络建立磨矿粒度预测模型，神经网络在这里充当黑匣子，通过模型就可以对选矿厂磨矿分级过程进行预测。”

“原来是这样”台下一学员沉吟着嘀咕道。

台下有人问道：“徐教授，这个神经网络方法进行预测的效果怎么样呢？”

徐教授回答道：“将训练好的磨矿粒度神经网络软测量模型用于某大型选矿厂磨矿过程进行磨矿粒度在线软测量。模型嵌于现场应用的磨矿过程优化软件中，通过监控机直接获取仪表的检测数据进行测量（测量结果与实验室化验结果的比较如下图所示）。仿真结果表明，该模型能够很好地实现磨矿粒度的在线测量，模型精度也满足工艺要求。”

最终，学员都明白了如何针对不同批次、品质的矿石实验确定最佳控制参数的原理和概要，以达到提高回收率和冶矿品位。



5.5　炼焦配煤优化

徐教授说道：“这节课我们讨论与冶金相关的另外一个数据挖掘的应用问题——炼焦配煤优化。大家知道，焦炭是冶金、机械、化工行业的主要原料和燃料。我国煤炭资源虽然比较丰富，但炼焦煤资源却相当贫乏，为了合理的利用煤炭资源，节约优质炼焦煤，生产出高质量的焦炭，亟需改进炼焦配煤技术。”

作为业内人士，玻璃公司的尚主任首先说道：“高炉用焦和铸造用焦要求比较高，大多数单种煤在焦炉内不易炼出机械强度较高的优质冶金焦。配煤过程涉及到把多种性质不同的单种煤，按照一定的比例进行配合，得到符合质量要求的配合煤。这种配合煤通过炼焦过程后，可以获得高炉炼铁用的焦炭。”

尚主任停顿了一会后，鼓风动力集团的王总说：“目前一般的焦化厂的炼焦配煤

工艺大致是：通过配煤槽，将各煤槽中的单种煤传送到输送皮带上，混合均匀后经过除铁和粉碎送往焦炉炼焦。”

徐教授补充说道：“在炼焦配煤过程中，存在确定配比时主观随意性大、准确性不高的问题。此外由于炼焦配煤控制系统的复杂性，传统 PID 控制方式对炼焦配煤过程的控制存在着控制精度不够高、实时性不够好等缺点。所以如何设计炼焦配煤优化模型，并实现其工业应用以期获得有利于炼焦配煤工艺的方法非常值得研究。”

听完前面的讲解后，工行的张行长问道：“徐教授，我是纯粹外行，一点也不懂这个炼焦配煤。—般而言，影响炼焦质量的主要因素有哪些呢？”

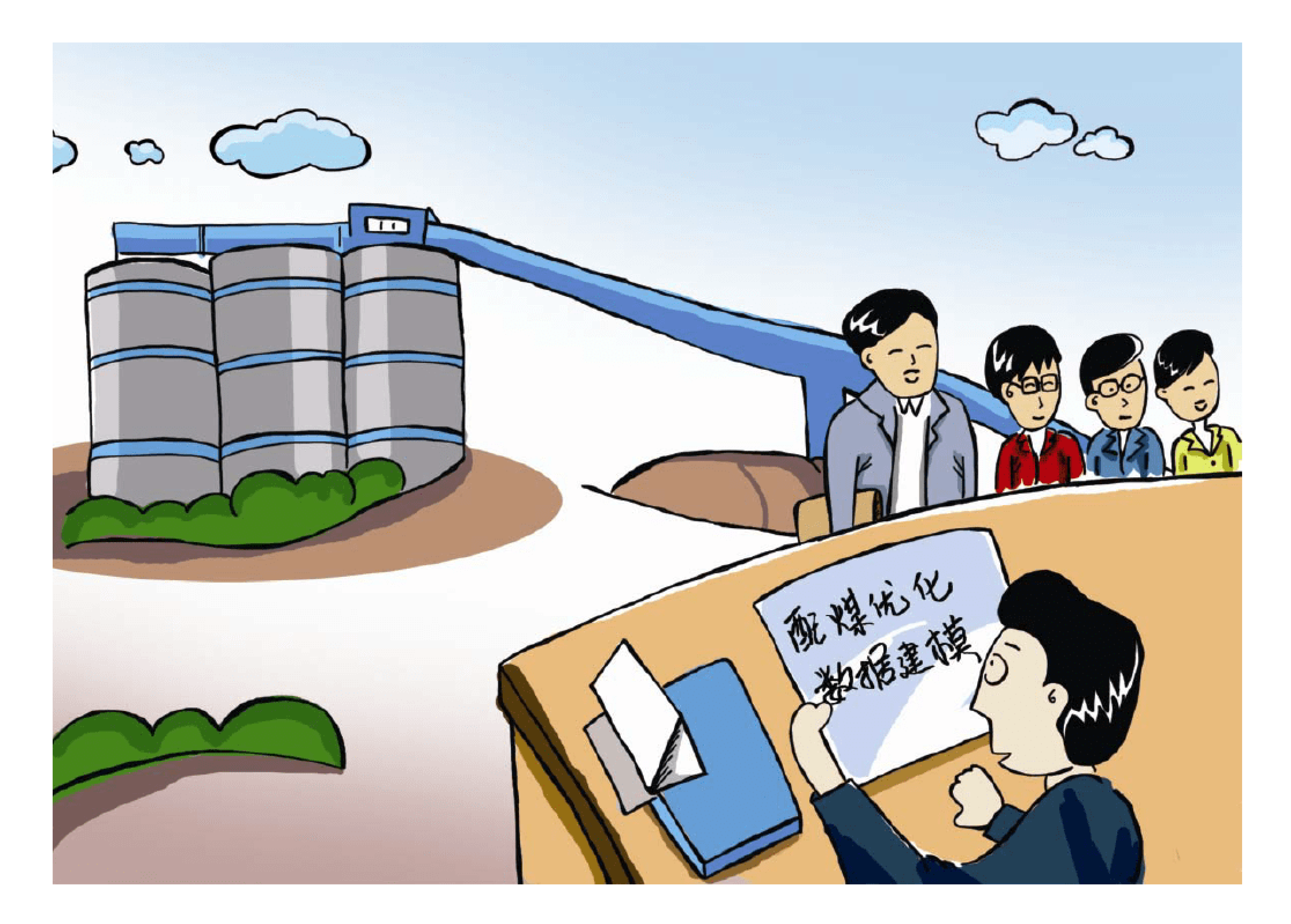
“我知道一个，水分。”刘经理抢先回答道。

徐教授说道：“确实是，回答的很好，值得表扬的一点是会用统计规律来判断。配合煤的含水量对焦炉的生产和焦炭质量都有很大影响。配煤水分高，炼焦耗热量大，结焦时间长，因而使焦炉生产能力降低。配合煤水分应力求稳定，以利于焦炉加热稳定，因此来煤应避免直接进配煤槽。”

“一般煤场等工厂治理污染的时候，都说要防止空气污染。比如防止产生酸雨，所以我猜测这个里面会不会有硫的成分呢？”李主任问道。

徐教授点评道：“李主任，你这个猜测很对，而且有推理，有神探柯南的风范。硫分对炼焦配煤的影响也比较大，焦炭含硫高，将使生铁含硫高，质量降低，同时还将影响炉渣的碱度。特别是含硫量波动较大时，对高炉操作指标的影响很大。”

电力公司的刘经理连忙说道：“上课之前我做了一点功课，学习了一点点工艺方面的常识。对炼焦影响较大的另外一个重要因素是灰分，在炼焦过程中，配合煤的灰分全部转入焦炭。灰分是硬度较大的惰性物质，配合煤灰分高，则粘结性较差，炼出的焦炭裂纹宽、深且长，强度低。”



李部长也信心满满地说道："在高炉冶炼中，高灰分的焦炭一方面在热作用下裂纹扩展，焦炭粉化影响透气性；另一方面，在高温下焦炭结构强度降低，热强度差，使焦炭在高炉内进一步被破坏，不能很好地起到骨架作用。降低配合煤灰分有利于降低焦炭灰分，可使高炉、化铁炉等降低焦耗，提高产量。"

"回答得很对，表现很好，提前做功课也非常值得表扬。这个课我越上越喜欢，开始享受这种上课方式。知道了这些关键因素之后，我们就可以通过对水分、灰分、硫成分控制整个炼焦配煤的工艺。"

"徐教授，你的意思是否这样理解：通过数据挖掘手段预测一下灰分、水分、硫分的比例，比如神经网络或者支撑向量机回归方法，这样就可以控制生产中的各参数，以实现最佳效果？"玻璃公司的尚主任提问道。

徐老师慈祥地笑道："真是一点就通，这里采取的主要手段是神经网络方法。研究表明，传统控制虽然取得了比较好的控制效果。但是，由于缺乏自学习能力，自适应能力差，使系统的鲁棒性受到限制。神经网络具有强大的自学习能力，可动态调整隶属函数，在线优化控制规则，设计出模糊神经网络控制器。应用在变参数的炼焦配煤系统对象模型，可以取得很好的控制效果。通过神经网络方法，我们可以建立焦炭水分、灰分、硫分的预测模型。"

"徐教授，那在建立模型时，应该选取哪些参数呢？"又有人问道。

徐教授说："拿灰分预测模型来说，建立预测模型时焦炭的灰分完全来自于炼焦煤的灰分，预测焦炭灰分的关键参数是炼焦煤灰分；关于硫分的预测模型，选择炼焦煤硫分和炼焦煤挥发硫分作为两个预测参数；关于水分的预测，稍微复杂一些，需要考虑相关的冷强度和热性质两个方面，焦炭冷强度的预测选择炼焦煤反射率和炼焦煤胶质层最大厚度等进行预测，焦炭热性质需要选择炼焦煤镜质组平均最大反射率、炼焦煤胶质层最大厚度、炼焦煤灰分、炼焦煤微强粘比、催化指数等。"

炼焦配煤模型	表征参数
硫分预测	炼焦煤硫分和炼焦煤挥发硫分
灰分预测	炼焦煤灰分
水分预测：冷强度	焦煤反射率、炼焦煤胶质层最大厚度
水分预测：热性质	平均最大反射率、炼焦煤胶质层最大厚度、炼焦煤灰分、炼焦煤微强粘比、催化指数

"徐教授，在利用神经网络建立上述预测模型时，需要把握哪些方面？"下面有学员提问道。

"首先，根据实际问题确定输入特征向量和隶属函数；其次，必须根据实际需要确定网络的拓扑结构，即网络具体由几层构成，每一层应该设置几个节点，合理的网络结构会使网络的学习收敛过程加快，有效减少网络的复杂性；第三，选择网络的算法，现在已有许多理论成熟的神经网络算法，每一种算法都有其优缺点，都有其适用的领域，因此，选择网络算法时要考虑到实际应用的需要及网络的推广与优化能力。"徐教授回答道。

“哦，原来这样就可以实现炼焦配煤模型优化设计了。”下面的学员恍然大悟。

“炼焦配煤工艺直接影响到焦炭生产的质量，针对传统的炼焦配煤工艺的影响因素，采用数据挖掘算法优化设计炼焦配煤工艺的控制模型，对于进一步提高炼焦配煤工艺的生产质量具有一定借鉴意义，值得大力推广。”

第 6 章　数据挖掘在税务、金融行业的应用

世界各地有许多国家每年都会因为纳税人的偷漏税问题而损失大量的财政收入。过去，税务稽查人员经常依靠以往的工作经验和某些直觉上的判断来圈定不法纳税人的特征。随着经济发展和税务体制的改革，税源、税种在不断增加，过程中累积了大量的税务数据。这时，以往的依靠经验和直觉判断区分违法纳税人的方式，势必会导致稽查成本增大、选案不科学、稽查效率低下等问题。借鉴国外的成功经验，使用数据挖掘，对税务管理部门所辖的纳税户进行纳税评估工作，建立动态、智能化的稽查选案，将会大大提高稽查工作的效果。本次 EMBA 将安排一节课来专门讲述**数据挖掘在税务稽查中的应用**。

由于科技的日新月异以及金融服务业的全球化，洗钱已经变得国际化和日益复杂化，成为一种“犯罪屏障”，让犯罪分子有恃无恐，危害甚大。世界各国从各方面加强反洗钱工作，比如完善制度建设及加强监管等，反洗钱获得了不少进展。值得一提的是，数据挖掘手段为洗钱客户的识别、大额可疑交易的及时发现并报告方面提供了一种可行的有力支撑。本次 EMBA 将安排一节课程讲述反**洗钱的内容**。

金融市场中，股票投资具有高风险与高利润并存的特点。科学的股票投资做法应该是根据预先设定的标准，选择股票，并对其在组合中的相对权重进行优化配置，使构建出的股票指数组合的追踪成本和风险控制在可接受范围内。针对股票市场进行市场追踪和收益优化，不论对金融机构还是个人投资者都意义重大。本次 EMBA 将安排一节课程来学习**数据挖掘在股票指数追踪中的应用**。

6.1 税务稽查

李部长今天来得特别早，一边翻看着手机报，一边注视着教室门口。看见孔部长走进教室，赶紧向他摆手，一边说道："老孔，来坐这，来坐这！"

孔部长冲李部长笑笑，和其他同学打了招呼就径直走到李部长旁边的位子坐下，问道："李大部长，有什么事情指教？"

李部长故意神秘兮兮的，趴到孔部长耳畔说："Google 偷税被查出来了，都上新闻了！你不知道？！"

"这个事，知道啊，又不是我开的 Google 公司，关我什么事？！"孔部长装作满不在乎。

"谷歌在中国大陆遭举报可能有逃税问题，北京市地税局第二稽查局正在对谷歌进行调查呢！亏你还是银行洗钱监察专家呢，这个都不关心?!"有点失落的李部长说。

"哈哈！我的李大部长啊，这个事情我昨天就知道了！这样的信息我怎么可能不关注呢？！你落伍啦！"孔部长说。



李部长被弄得摸不着头绪了，说：“不可能！我今天才看到的新闻啊！”

“确实是昨天的新闻啊！老李啊！”孔部长大笑着说。

李部长仔细一看，原来自己昨天的手机报没看，现在看的正是昨天的手机报，尴尬地说：“唉！我今天看的是昨天的新闻，其实我就是想幽默下，哈哈！”李部长给自己找了个台阶下，接着说：“老孔，现在税务稽查那么严格，谷歌是怎么逃税的？”

“这个还真不太清楚。”孔部长说。

“说什么呢？这么热闹！”李部长和孔部长说得正起劲，却没发现徐教授已到了身旁。

孔部长发自内心的仰慕，对徐教授说："刚说谷歌逃税呢，正有问题想请教徐老您呢！"

"是吗？今天我们的话题就是数据挖掘在税务行业的应用，不妨我们大家一起来好好讨论一下。"徐教授说。

说完徐教授径直走向讲台，说道："大家静一下，开始上课了！"大家都把手机调成震动，打开笔记本准备听课。

徐教授说："这两天发生一件新闻，相信大家都听说了，搜索引擎公司谷歌涉嫌逃税。"

"Google 怎么可能逃税呢？"有人问道。

税务局姚局长开了口："Google 从 2000 年开始向中国网民提供中文搜索服务，2003 年再推出中文关键词广告。在这段期间，中国客户只要拥有一张国际信用卡，把钱直接打入 Google 在美国的账号，就可以在 Google 网上购买关键词广告。虽然发布广告的客户和由此产生的点击收入都来自中国，但是相关的资金流转在中国却没有任何记录。作为 Google 曾经的客户都没有从 Google 拿到正式发票，客户虽然通过代理商可以得到正式发票，但开票单位是代理商而非 Google，是代理商和 Google 间的结算。"

"哦，原来这样！真够狡猾的！"台下的学员听姚局长说了内幕，恍然大悟。

"那还有一个问题，是怎么发现谷歌逃税的？"李部长问道。

"据说有人举报才发现的。"姚局长回答道。

孔部长听完笑着说："光靠举报恐怕不行吧！目前税务部门是如何来进行稽查选案的呢？"

姚局长说："目前我国税务机关主要采用人工选案的方法。人工选案主要是税务稽查工作人员根据纳税人名单，按企业性质、按企业工作方式、按企业规模等方式进行排查。"

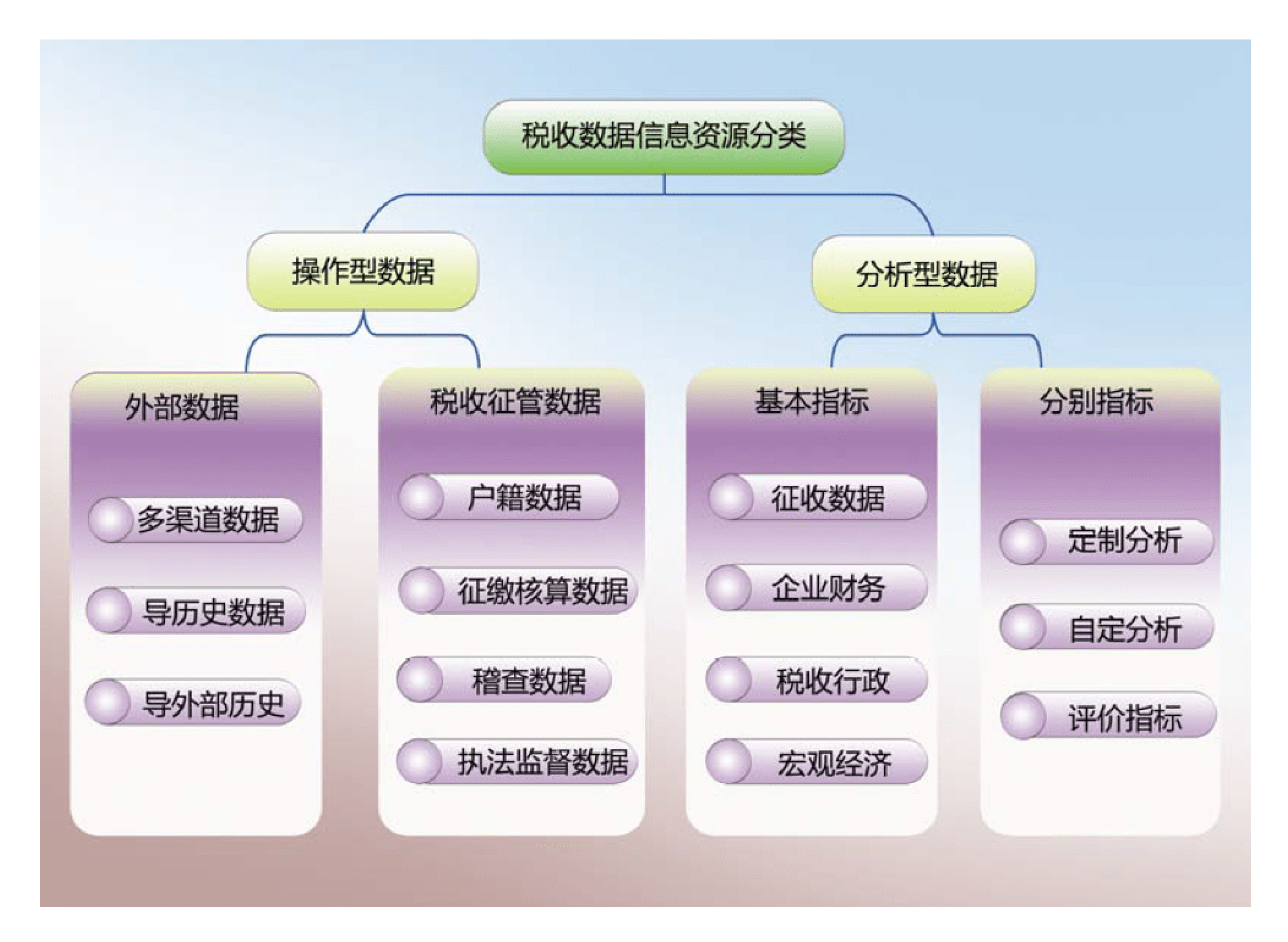
"这种方法效果怎么样？"李部长侧着耳朵听了好久，转过头问姚局长。

姚局长声音洪亮地说："这种方法存在着很大的盲目性和随意性，至于哪些企业有问题，哪些企业需检查，经常是按照以往稽查经验或者主观臆断来进行稽查，结果导致稽查成功率很低并且造成税务机关人力、物力、财力的浪费，也增加了纳税人的负担。随着经济的发展，自然而然地引起税源、税种的增加，这时仍然沿用以前的老方法，一方面大面积撒网会使稽查成本增大，选案的不科学性也会引起稽查效率低下，甚至会有较多的漏网之鱼。"

徐教授看着台下的学生，意味深长地说道："我国已经成为世界第二大经济实体国，提高税务稽查工作的效率，已成为当务之急。税务机关必须把现代化的科学手段引进税务稽查中，实行计算机智能选案，也就是利用数据挖掘技术，将稽查经验模型化，然后对纳税资料进行系统地分析、对比、排列和组合，从中列出税务稽查重点对象。"

"数据挖掘首先需要有充分的数据积累，现在税务部门这方面的条件具备吗？"有人担心地问道。

税务局姚局长回应道："应该没有问题。2001 年起全国税务系统形成了以增值税发票交叉稽核比对为主要内容的'金税工程'广域网络系统，积累的数据量激增；之后又推行了统一的征管系统，建设了市一级的集中数据处理中心，市、县、乡镇基层分所的业务统一到了市局，实现了集中管理，现在正在推行省一级的集中数据处理模式，税务数据累积的速度和数量增长更是惊人。"



徐教授说："随着数据的不断积累，数据挖掘在税务行业将更有用武之地。"

"发达国家税务稽查选案有什么经验可以借鉴？"李部长问道

"美国的数据挖掘技术应用于税务行业的时间比较早，90%以上的案件是通过计算机程序对纳税信息的分析而筛选出来的，基层稽查人员只有不到 10%的参与决策权。"姚局长介绍国外的情况。

"啧啧！瞧人家这水平！"有人慨叹不已。

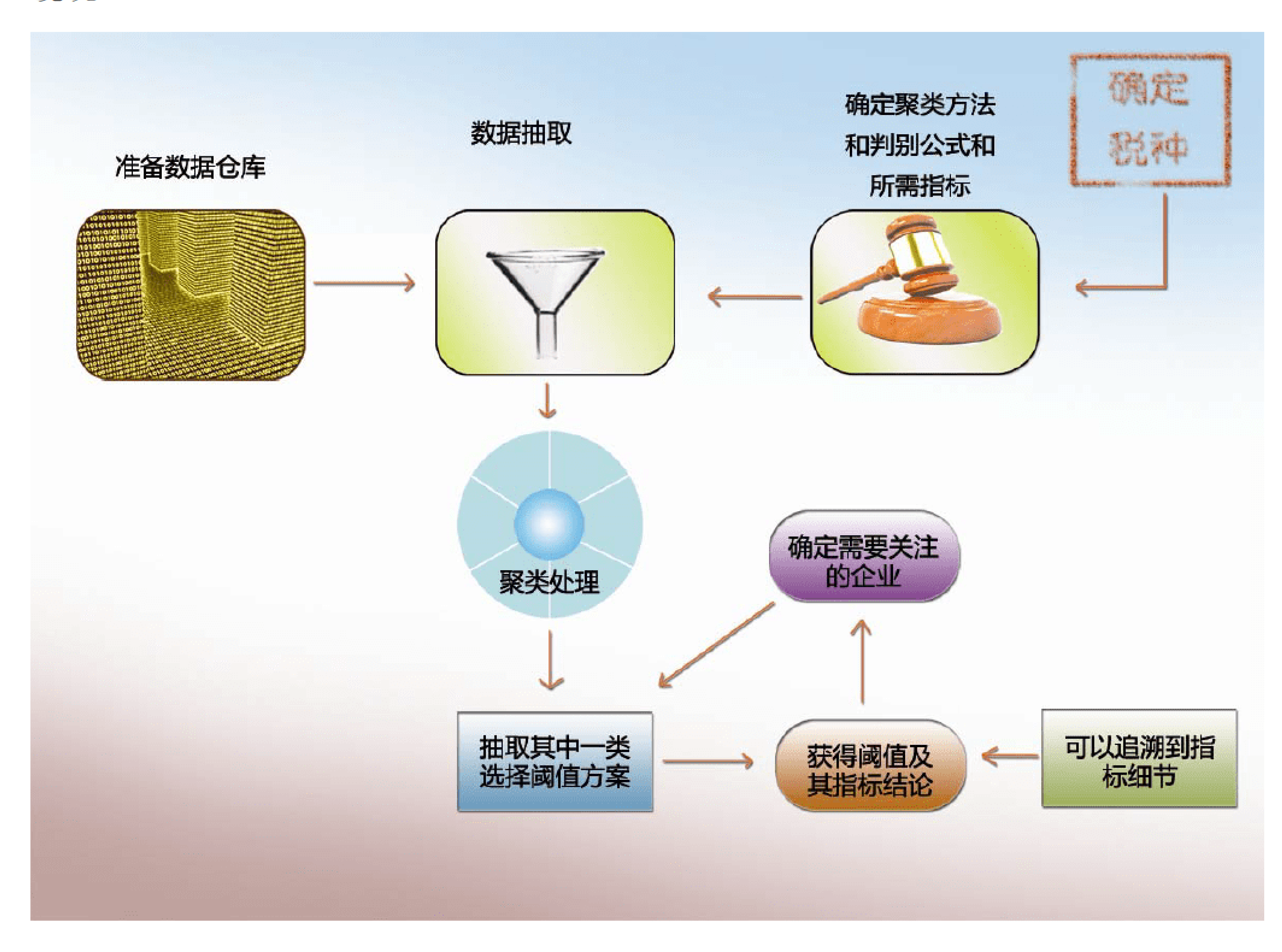
"有了国外经验，咱们就可以少走弯路了哦！"台下有人说。

徐教授说："是这么个道理！从国外税务稽查的经验看，科学技术力量的充分运用，财力、物力、人力的优化配置以及强大执法权力的法律保障是做好税务稽查工作的基础。案源信息的完备（如纳税人各种基础资料的收集与评估的准确和完整性，法

定的广泛信息共享及获取权和强大的信息分析及评估利用）为税务稽查科学化、精细化提供了有力保证。”

“徐老师，要利用数据挖掘技术实现计算机智能选案，具体应该怎么做呢？”

“首先我们要选取税收有关指标，然后用聚类分析方法进行聚类，再对每一个簇选取有关指标，对指标进行遍历，发现异常进行标记，如果遍历结束都没有发生异常，很好，我们就认为那一簇的企业都是守法单位。请看大屏幕。”徐教授向上托了托眼镜说。



“原来是这样，不错！很强大！偷税漏税无处逃了。”

“这只是举个简单的实施例子，其实如果实施起来中间会有很多复杂的细节。为了更好地提高税务行业的效率、公平，为国家严把税务大门，这样一个简单系统还是不够的，我们需要建立基于数据挖掘技术的税务决策支持系统。”徐教授说。

6.2 反洗钱

徐教授："2008 年年初，陈水扁家在瑞士的 2100 万美元遭冻结曝光，由此牵扯出陈水扁家族将巨额款项洗钱到海外。2008 年 8 月 14 日陈水扁成为阶下囚。本节课，就让我们一起来研究一下洗钱的相关问题。谁先来告诉我们什么是洗钱？"



工行的张行长回答说："洗钱指将毒品犯罪、黑社会性质的组织犯罪、恐怖活动犯罪、走私犯罪或者其他违法所得及其产生的收益，通过各种手段掩饰、隐瞒其来源和性质，使其在形式上合法化的行为。"

"回答的非常专业。洗钱活动最早出现在 20 世纪 20 年代，当时美国芝加哥的一名黑手党成员开了一家洗衣店，在每晚计算当天的洗衣收入时，他把非法收入混入洗衣收入中，再向税务部门纳税，扣去应缴的税款后，剩下的非法所得就成了他的合法

收入。一般地，黑钱的非法来源途径有：贩毒、走私、贩卖军火、诈骗、盗窃、抢劫、贪污、偷税漏税等犯罪活动”，徐教授说道。

“徐教授，那这些黑钱是怎么转化成合法的金钱的呢？”下面一个学员问道。

徐教授回答说：“洗钱主要手段有以下几种。第一种，也是最容易想到的方法：存进银行，以本人、亲属或者其他人名义，甚至用化名或假名，将非法所得存入银行，变成合法存款的一部分；第二种，搞‘一家两制’，一边非法捞钱，一边授意亲属或子女创办现金密集型的公司或企业，将非法所得混入营业收入一并申报纳税，以掩饰非法资金的真实来源；第三种，通过‘稻草人’打理开办的公司，企业表面上是他人的，实际上由自己控制；第四种，通过地下钱庄、赌场等将黑钱转移出国出境，也有些是直接安排在境外收取赃款并就地清洗。此外，新的洗钱方式是在科技进步和金融创新的形势下出现的，如通过通信账户、网银、国际互联网银行、智能卡进行洗钱等。”



“手段还真多样化，难怪洗钱演化的这么猖獗。这些人真是老百姓和国家的祸害，应该严厉打击！”台下马处长气愤地说。

工行的张行长接过话题说道：“就是，洗钱不仅造成了极其严重的经济、安全和社会后果，同时还为贩毒者、恐怖主义分子、非法武器交易商、腐败的政府官员以及其他罪犯的运作和发展提供了动力。”

徐教授说道：“正是这样，反洗钱行为应运而生。各国政府动用司法力量，比如颁布《反洗钱法》，调动有关的组织和商业机构对可能的洗钱活动予以识别，对相关机构和人士予以惩罚，从而达到阻止洗钱犯罪活动目的。”

工行的张行长说道：“洗钱已经变得越来越国际化，而与犯罪活动有关的金融问题也由于科技的日新月异以及金融服务业的全球化而变得日益复杂化。所以如何进行反洗钱、有效地甄别出犯罪特征、与犯罪份子作斗争也越来越难”。

徐教授补充道：“以 911 事件为例，乘坐美国航空公司的恐怖分子持学生签证进入美国，他们的银行账户有大量资金进出，而且大多数是从已知的、支持恐怖主义的国家大笔电汇的，但几乎没有典型的学生消费支出。”

“如果在炸五角大楼之前，美国安全局接收到关于银行的可疑洗钱报告就好了，能避免一场灾难。”下面有人感慨道。

“事实是收到了那个关于洗钱的可疑报告，但是结果也不会有什么改变。因为每年各金融机构提交的可疑报告数非常多，可最后根据线索确定是洗钱的比例占可疑报告总数不到万分之六，所以没有引起足够重视。”徐教授说道。

马处长感慨地说：“感觉这有点类似于中国古代寓言故事——狼来了，刚开始的时候监管部门会比较重视验证真实性。随着可疑交易报告数量的增多，报告的边际信息价值递减，对真正可疑交易行为的发现几率趋于降低。等到最后狼真的来了时，人们已经对事情置若罔闻了，所以注定要成为狼的盘中之餐。”

“徐教授，我对这个问题比较好奇：当时的美国银行是根据什么方法判断出来那些恐怖分子的可疑行为的呢？”马处长问。

徐教授："从以交易为导向的数据观转变为以对象（人或组织）为导向的数据观，美国的 FAIS 系统综合使用了人工智能技术和基于案例的推理、黑板等技术制定出来了 336 条规则。美国银行的每一笔交易、每一个对象、每一个账户都要用 336 条规则去测试，每条规则都对这些交易、对象、账户给出非法或合法的判定依据，最后对每一个项目的可疑性进行评定。"

"徐教授，你刚讲在 911 事件中恐怖分子以学生身份进入美国，但是他们的银行行为记录不符合学生的特征。我想问的是，在甄别洗钱犯罪行为时，对这个银行的转账等特征提取是不是非常关键？"马处长道出了自己的困惑。

"这个问题问得非常好。洗钱者煞费苦心地掩饰其资金的非法性质和来源，使之混同于合法资金，并模糊稽核和审计的线索，他们不会优先考虑成本和效益，也不以追求利润最大化和节省费用为目标。洗钱是一种非理性的经济活动，因而必然表现出不同于正常理性的经济活动的征。"顿了顿后，徐教授接着说道："可疑洗钱行为在金额维度上表现为交易金额异常增大和近似等额两个特点，在时间维度上则表现为交易频率的异常变化。当账户被洗钱分子用于洗钱目的时，势必要增加交易的频次。虽然这些交易变化对于账户而言并不一定是不正常的，但是，如果这些交易的频次偏离了账户的正常交易行为模式，就值得引起我们的注意。"

"徐教授，要了解这些账户消费行为以洞察洗钱犯罪活动，需要知道哪些数据有助于我们进行反洗钱活动呢？除了你刚介绍的交易金额、交易频率等数据之外。"台下一个学员问道。

"根据不同的数据挖掘思路和方法，需要的数据有一定的差别。总的来说，有一些数据是都需要的，比如前面我们说的交易金额、交易次数。此外账户日常交易的信息，比如账号、交易时间、交易名称、公司名称、企业行业代码、企业性质等字段都可以作为可疑洗钱行为模式识别的研究属性。此外，在有些方法中还会考虑一些其他字段，比如上述交易行为字段再构造出的新变量、账户企业的信用等级、注册资金等账户属性，也可以用来研究可疑洗钱行为模式识别。比如有研究表明，对企业类客户

而言，其交易可疑与否与交易次数和交易金额直接相关之外，同时还与企业的经济性质、注册资金和信用等级相关。”徐教授解释道。

“徐教授，我知道有了这些数据之后，必须进行一些数据预处理工作来防止‘垃圾进垃圾出’，以保证数据的质量。比如数据中的重复性信息、缺失值的替换、源数据的分布特征等探索性工作。在进行完数据前期准备工作之后，就可以用具体的数据挖掘手段来解决问题了。我的问题出来了，在反洗钱活动中可以使用的数据挖掘方法有哪些呢？”下面一个学员问道。

“呵呵，今天的这个课上着上着越像‘你问我答’了。关于这个问题，结合前面的学习，有没有谁先表达一下自己的意见？”徐教授笑着说道。

这时，马处长站起来说道：“那我就先说一下：根据已知可疑行为模式和不可疑行为模式的历史信息，运用回归分析等预测技术来建立预测模型，计算任何新进入账户的行为可疑概率。这是我自己的一点浅薄意见，可能不是很成熟，大家见笑了。”

徐教授点评道：“回答的不错，是个解决思路。但是，预测模型要求用已知洗钱和非洗钱行为模式的历史数据作为训练样本，但这正是许多机构所欠缺的。还有谁想分享一下自己的看法？”

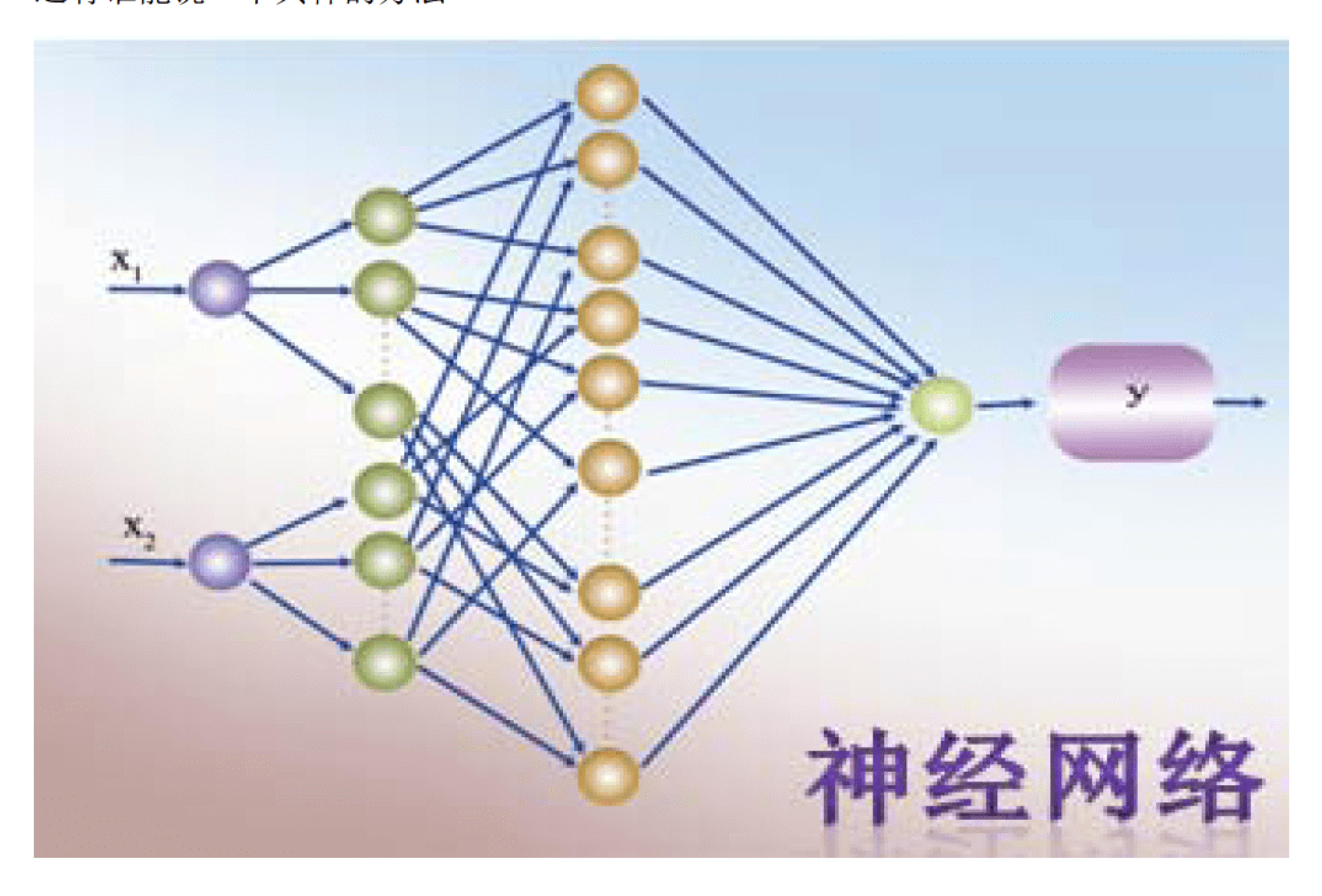
徐教授刚说完，李经理便自告奋勇地站起来说道：“前面我记得徐教授讲过一句话，说神经网络可以用来预测也可以用来分类。所以我想是不是可以将银行账号通过神经网络进行分类呢？将所有人贴上标签，分为正常的和可疑的两种。”

之后，徐教授点评地说道：“回答的不错，对于不完整信息、错误、不准确信息的高维数据集，神经网络毫无疑问成为了有用的、抗噪声干扰的统计模型，具有显著的拟合观测数据的能力。”

听完徐教授的介绍，李经理锲而不舍地追问道：“具体通过神经网络怎么实现我还不晓得，看徐教授能不能通过一种神经网络方法介绍一下。”

徐教授回答道：“那就给大家说一种神经网络方法：径向基神经网络。在径向基神经网络中仅包含一个隐含层，隐层神经元采用径向基函数作为其输出特征，径向基

是一种非线性映射，输入层到隐层的权重值均固定为 1，输出节点为线性求和的神经元，隐层到输出层的节点之间的权重可以调节，输出层为隐层节点的加权求和。总之，径向基神经网络就是将原始的线性不可分的特征空间变换到另一个空间，通过合理选择这个变换，使得在新空间中原问题线性可分，最后利用这个线性神经元来解决问题。根据数据分块的调整、多次隐层节点的变化尝试以及各种不同的可变参数设置，利用模型的误分率来选择合适的神经网络方法。遗憾的是不能给出明确的函数结构或者规则，只能将整个神经网络建模流程作为黑匣子来使用。在反洗钱的数据挖掘应用中，还有谁能说一下具体的方法？”



“既然李经理刚才说的神经网络进行洗钱犯罪行为侦查，采用的是分类技术。那前面学习的决策树也可以用来分类，根据历史数据，为每个数据对象加上分类标签，使其成为训练数据集，并采集与其相关的属性值，选择一个启发式规则或统计度量，如信息增益或基尼系数，将训练集进行反复分叉训练，直至分叉后的训练集类别与事实分类一致为止。经过训练后建立分类模型，就可以对新进入的其他数据进行分类，

以判断其是否可疑。所以按照这个推理，决策树也是可以用在洗钱侦查识别中的。”台下的章主任也信心满满地说道。

“呵呵，逻辑推理能力非常好。还有人有不同的看法没？”

“徐教授，洗钱的话，这个转账方和接受转账方之间的交易行为应该在一定时期内是有一定固定性的。比如贪污的钱汇给自己在国外儿子（或者假名）的账户等。能否通过观察这些交易链来打击洗钱活动呢？”

“这个思路很独特，非常好。你说的这个技术叫链接分析，对反洗钱金融大额交易数据进行分析，从而找到有交易链接关系的可疑金融交易数据。为了得到更好的数据挖掘模型，对源数据进行探索性分析，了解基本分布特征。通过链接分析，交易双方作为连接节点，以交易先后时间确定资金流动方向，根据交易频繁度，确定异常的交易链。”

经过徐教授的肯定，大家都对前面自己学习的效果有了一定把握，感觉到自身的收获，都非常高兴。

徐教授也高兴地说道：“有了这个武器，必然会给洗钱犯罪活动投下一枚炸弹。下面我再说一种方法：聚类。聚类是将研究对象的集合进行分组，形成由类似对象组成多个类别的过程。在研究中，划分到同一类的对象就是同类，没有分到同一类的就是异类。在可疑洗钱行为模式识别中，被分到不同类的也可能是同类，那些分类中的孤立点才是真正的研究对象。在大多数情况下，孤立点的判断标准是隐含的，不能轻易地从聚类过程中推导出来。谨慎选择警兆指标，使用基于距离的聚类算法和基于网格的聚类算法来识别可疑洗钱行为模式，这样就能区分出正常与可疑，而不是简单地区别正常与异常。”

“徐教授，这里你提到两种聚类方法：基于距离的和基于网格的。这两种方法在实际应用中各有什么优势呢？”

“基于距离的聚类方法优势在于该算法不需要预先设定簇的个数、比较容易实现、时间复杂度较低且可以处理海量交易数据，可以用来解决偶然行为识别；而对惯

常可疑行为模式的识别，则选取具有较强包容性的网格聚类算法，因为它不需要预先设定簇的数量、能发现任何形状的簇、不受噪声影响。此外，基于网格的聚类方法能从分析结果中过滤处于稠密区域的大量主体数据，只以剩下的高离群度的数据作为基础挖掘数据，可以有效地减少计算量，提高计算效率。”

“结合徐教授的聚类方法，现在我们的反洗钱数据挖掘技术的能量肯定是导弹级别了”，刘经理幽默地说完后，台下学员也跟着笑了起来。



“徐教授，我还有最后一个问题。通过刚才讨论学习，我们已经知道建立分类模型的方法了，比如神经网络、决策树等。那我们怎么检验这个模型就是合理的呢？”台下有人问道。

徐教授回答说：“这就涉及到模型的评估问题，针对模型进行评估，有很多指标。比如可以检测模型目标变量的提升曲线，如果提升曲线是递减的说明模型是有效的，在有效的前提下，提升率越高的曲线代表了模型的拟合度越高。”

在本节课的最后，徐教授强调：“任何一种方法都有一定的使用范围和局限性。在反洗钱中应用数据挖掘技术，我们更倾向于找出可以发现刻意交易相关信息的方法，不在于给出一个绝对可以使用的结果。”

6.3　股票指数追踪

今天马处长和徐教授在路上一直在讨论，不时听到他们爽朗的笑声。好奇心比较重的李部长看见他们进了教室，赶快摆手招呼马处长过来。

马处长微笑着走过去，拍下李部长的肩膀说：“老李，又看昨天新闻报了？是不是谷歌又逃税了？”

因为上次李部长看‘过期手机报’事件而受到马处长的取笑，他并没有感觉尴尬。而是幽默地说：“我又‘温故而知新’一次，哈哈，老马你和徐教授在路上讨论什么呢？”

马处长故意装神秘地说：“秘密！”

看到李部长胃口吊起来了，马处长依旧用神秘的语气说：“指数追踪！”

由于马处长声音小，李部长没听清楚，说：“碧海追踪？！”

李部长的声音可不小，惹得大家一阵大笑，被讲台上徐教授听到了，说：“我和马处长可没看电影‘碧海追踪’，我们讨论的是股票指数追踪！”

“指数追踪？你俩追踪谁？”李部长丈二和尚摸不着头脑，继续问。

“我们俩讨论用数据挖掘技术进行股票指数追踪呢，没有干什么不正当的‘追踪勾当’，哈哈！”

“指数追踪嘛，我知道，很有意思哦！”李部长装作知道指数追踪，给自己找了台阶下。

“有谁能说下指数追踪的概念？”徐教授问。

“指数追踪是指用资本市场上若干个金融资产的组合来追踪市场上某一指数的表现。” 马处长把刚在路上跟徐教授学的东西说给了大家。

“指数追踪的原理是怎样的？”有人问。

“指数追踪是指通过利用一个股票组合复制某一现实指数或者虚拟指数的市场表现，来获取与指数相近的收益，试图最小化跟踪误差。通常来说，一般的指数追踪技术关注于最小化跟踪误差的方差，并考虑组合收益与标的指数收益的相关性，或者是组合调整的交易成本最小化。”

“哦，这么说指数追踪是比较困难的喽！”台下有人说。

“市场上的股票指数往往包含几百上千种股票，即便是以市场指数为参考，想要以有限资金按照股票指数的构成比例购买所有的股票，来追踪其波动的确是非常困难的。”

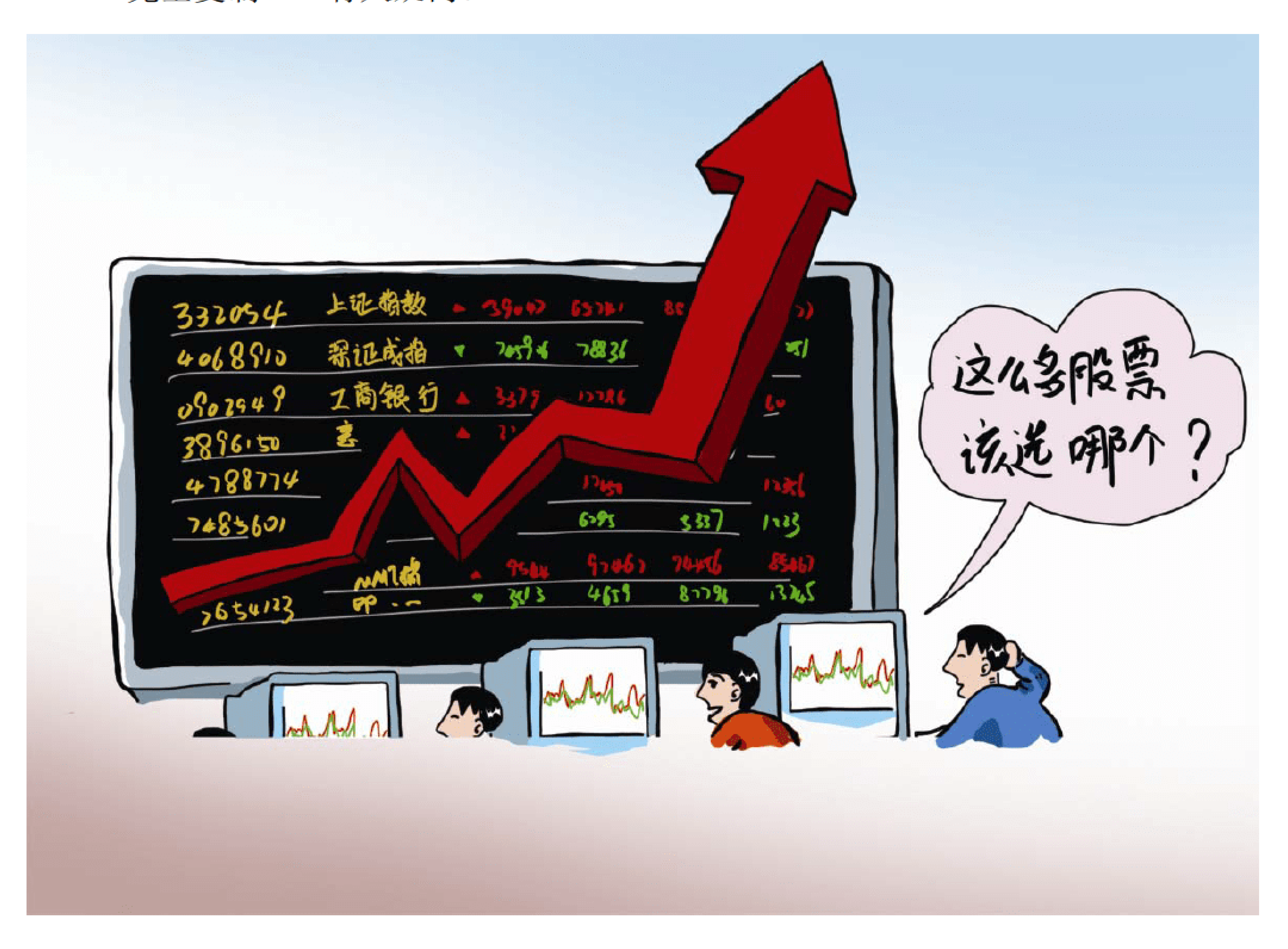
“有什么方法来解决这个问题呢？”李部长问。

“那下面就要提下指数组合优化了！指数组合优化是通过权重的优化再配置来寻找一个含有‘部分’成分证券的最优的追踪组合，所谓的‘最优’就是使得该组合相对标的指数的追踪误差最小或其他事先设定的标准最优。其目的在于复制与该指数同样收益水平的一个投资组合，实现组合收益与指数涨跌基本一致。”徐教授说。

“哦，原来是这样！”

“因此，研究具有高精度、低交易成本、且能保证追踪组合的高流动性的指数追踪技术有重要的意义。目前，指数追踪技术主要分为两大类，即完全复制和不完全复制。”徐教授接着讲。

“完全复制？”有人疑问。



徐教授解释说：“完全复制是指通过按照目标指数成分证券在目标指数所占权重来购买所有成分证券，构建追踪组合对指数进行追踪，这种方法由于成本高，管理复杂，一般很少用到。”

“那么来说，完全复制具有很高的精度喽！”孔部长说。

徐教授说：“是的，从理论上讲，如果采用完全复制的策略应该不会存在追踪误差，但是实际中并非如此，例如，当标的指数的构成发生变化时，该指数假设所有股

票在理论组合中的权重能够自动实现。然而，指数基金经理并不能这样假设，他们需要对股票的权重进行现实调整以达到模拟指数的目的。”

“那不完全复制肯定就是从证券样本中抽取一部分来进行分析了。”李部长凭借自己聪慧的脑瓜在大家面前闪光了一下。

徐教授说：“不错！不完全复制是指通过利用所有成分证券的子集中包含的证券按照一定比例构成的组合来追踪指数，包括优化复制和分层抽样复制。”

“优化复制？这么多新词啊！”

徐教授接着说：“优化复制是直接用优化方法确定进入追踪组合内的成分证券及其投资权重，而分层抽样复制则是先按行业、流通市值、换手率等指标人为确定进入追踪组合的证券，再通过优化方法来确定各成分证券的权重，从而有效改善追踪误差，提高追踪精度。依据经验的分层抽样确定的追踪组合在样本内外追踪效果并不一定很好，因此，优化复制技术得到了研究者和实际工作者的青睐。”

“这个可以理解。”

“对指数追踪的研究尽管丰富，但大多数研究角度都是基于样本内追踪误差最小，然后假设市场是有效的，因此认为依据样本内经验风险最小构建的追踪组合在样本外的追踪误差也是最小的。”

“指数追踪具体都有哪些方法？”李部长再次扮演了一个爱问问题的角色。

徐教授说：“指数追踪技术大致可以分为四种，首先是基于优化方法的指数追踪技术，第二个是基于协整的经典指数追踪，第三个是基于协整的增强型指数追踪，第四种就是基于协整的多头/空头统计套利策略。”

“那您给讲个较为简单而且常见的指数追踪技术吧。”孔部长一边详细做笔记，一边说。

徐教授看到大家对知识的渴望，说："好，通常的指数追踪技术一般采用优化方法，最为常见的是 TEV（追踪误差方差）最小化模型。可以用大屏幕显示的数学公式来表达，请看大屏幕。"

$$r_{index,t} = \sum_{k=1}^{n} c_k r_{k,t} + e_t$$

徐教授解释说："其中，左端的 r 是在 t 时刻的指数对数收益率，而是 k 股票在 t 时刻的对数收益率，是持仓权重，而代表追踪误差，一般的优化方法就是在约束条件——跟踪误差期望等于 0 和股票权重和等于 1 的条件下，利用数值方法使得跟踪误差的方差最小化。"

"不错，这个方法比较容易理解。"

"但是这个方法还是存在诸多不足的，例如该优化方法在被动投资中的缺点比较显著，首先，股票指数是组合内股票的一个线性组合，针对股票指数追踪误差最小化的过程中包含了许多噪音，依赖于样本数据。"

"哦，是的，这个指数追踪方法应该在高波动的市场中极不稳定。"马处长补充说。

徐教授继续说："其次，由于采用了相关系数来衡量协同波动，存在以下不足：首先，只能用平稳数据，如股票收益率，由于股票价格的差分序列损失了一些有用信息；其次，这只是一个短期的统计量，缺乏稳定性；第三，依赖于估计模型，相关系数易于受到异常值、非平稳序列或是波动率聚集的影响，因此在长期时间序列中可能会得出错误的结论。"

"针对这个方法存在这样或者那样的问题，还有没有其他较好的方法？"马处长问道。

徐教授回答说："比如有人研究提出了基于 L1 正则化的优化复制技术来实现最佳不完全复制的指数追踪问题，并应用到实际中去。"

这时，证券基金的方科长说道：“这个方法我们部门的人用过，这个方法也是采用优化抽样复制技术，比如从 N 支标的指数的成分股票中选出 k 支股票构建一个投资组合，并使得该组合相对标的指数的一些考核标准最优。”

方科长顿了顿，接着说：“但是，基于 L1 理论发展出的指数追踪方法选取的股票支数还比较多，投资管理中难以操作。”

徐教授解释道：“果然是内行，讲的知识点都比较细致。你说得很有道理，为应对这个缺陷，我们独特性地提出 L1/2 正则化的稀疏指数跟踪模型，在不降低追踪精度的情况下，选取的股票支数更少。”

方科长接着说：“徐老师，L1/2 正则化模型我可是头一回听说。对比 L1 正则化方法解决指数追踪问题，L1/2 正则化有什么优势呢？”

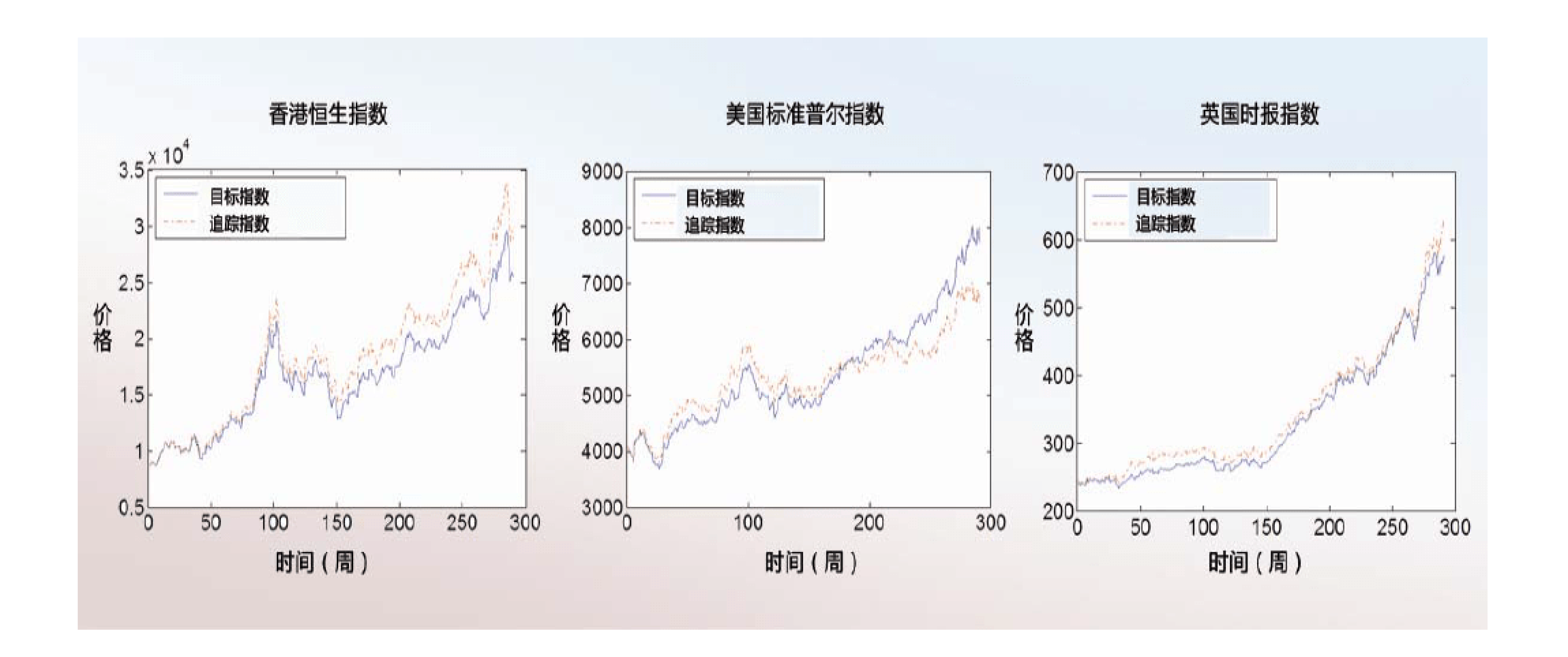
徐教授回答道：“最近的研究表明，Lq（$0<q<1$）正则化问题是非凸的、非光滑的、难解的，q 越小，其解越稀疏，但当 $q<1/2$ 后，稀疏性改变地不太明显了，也就是说 L1/2 是 Lq（$0<q<1$）的典型代表。而且对 L1/2 正则化模型，我们还构造了非常强大的迭代阈值的算法。它可以用最少的股票来追踪目标指数变化，它比 L1 正则化方法具有更好的稀疏性。”

这时，台下的马处长激动地拍了一下手，站起来说道：“徐教授，我想起来了。之前看一个新闻说您参加 2010 国际数学家大会，所做的报告中重要的一部分就是 L1/2 正则化理论。”

看徐教授笑着默认了，大家都很激动。

有学员站起来说道：“徐教授，您真是太让我们敬仰了。能研究出这么独创性的理论，我相信，这个 L1/2 正则化方法应用在股票指数追踪中效果肯定是非常好的。”

徐教授谦虚地说道：“实践是检验真理的唯一标准。针对香港恒生指数、英国时报指数、美国标准普尔指数等股票组合进行了成功的指数跟踪测试。试验表明，L1/2 模型可以有效地解决指数跟踪问题，保证较高的预测跟踪性。”


香港恒生指数
美国标准普尔指数
英国时报指数
目标指数
追踪指数
价格
时间（周）
x 10^4

第 7 章　数据挖掘在故障诊断中的应用

本节课一开始，徐教授上讲台便开口道：“先给大伙讲个故事。话说某厂的大型电机坏了，厂内技术人员都不知道毛病出在哪里。怎么办呀？”

“请外援高手吧！”台下一个学员出了个主意。



徐教授接着讲：“厂长无奈只好联系生产厂家，对方推荐了一个有经验的老工程师。工程师说要一千美元，厂长正着急，就答应了。工程师用仪表测了一会儿，然后拿起粉笔在电机的某位置画了一个圈圈，说问题就在这里，最后证明确实如此……”

“这年头画圆圈的都是高人，邓小平也是在南方画了一个圈。”台下一个人插话道。

徐教授笑着说：“设备修好了，工程师找厂长要钱。厂长看他只是那么轻松地画了个圆圈就要拿走一千美元，感到实在舍不得。但是又不好反悔，就让工程师列个维修清单出来，想难为他，迫使他降价。”

“列什么清单，纯技术活，没有维修材料，怎么列，就是想赖账！”有人愤愤道。

徐教授说：“工程师的维修清单：（1）用粉笔画圆圈，1 美元；（2）知道在哪里画，999 美元。”

听完徐教授的话，大家都笑了。

“果然够聪明，难怪能成为大牛。”台下的学员感叹道。

于是，徐教授趁热打铁说：“今天要给大家讲的内容就是如何利用数据挖掘技术，进行故障诊断，为企业少花这 999 美元。”

7.1　火箭发动机故障诊断

徐教授直奔主题，说：“在讲火箭发动机故障诊断之前，我先问大家一个问题。在中国被称为飞天第一人的是谁？”

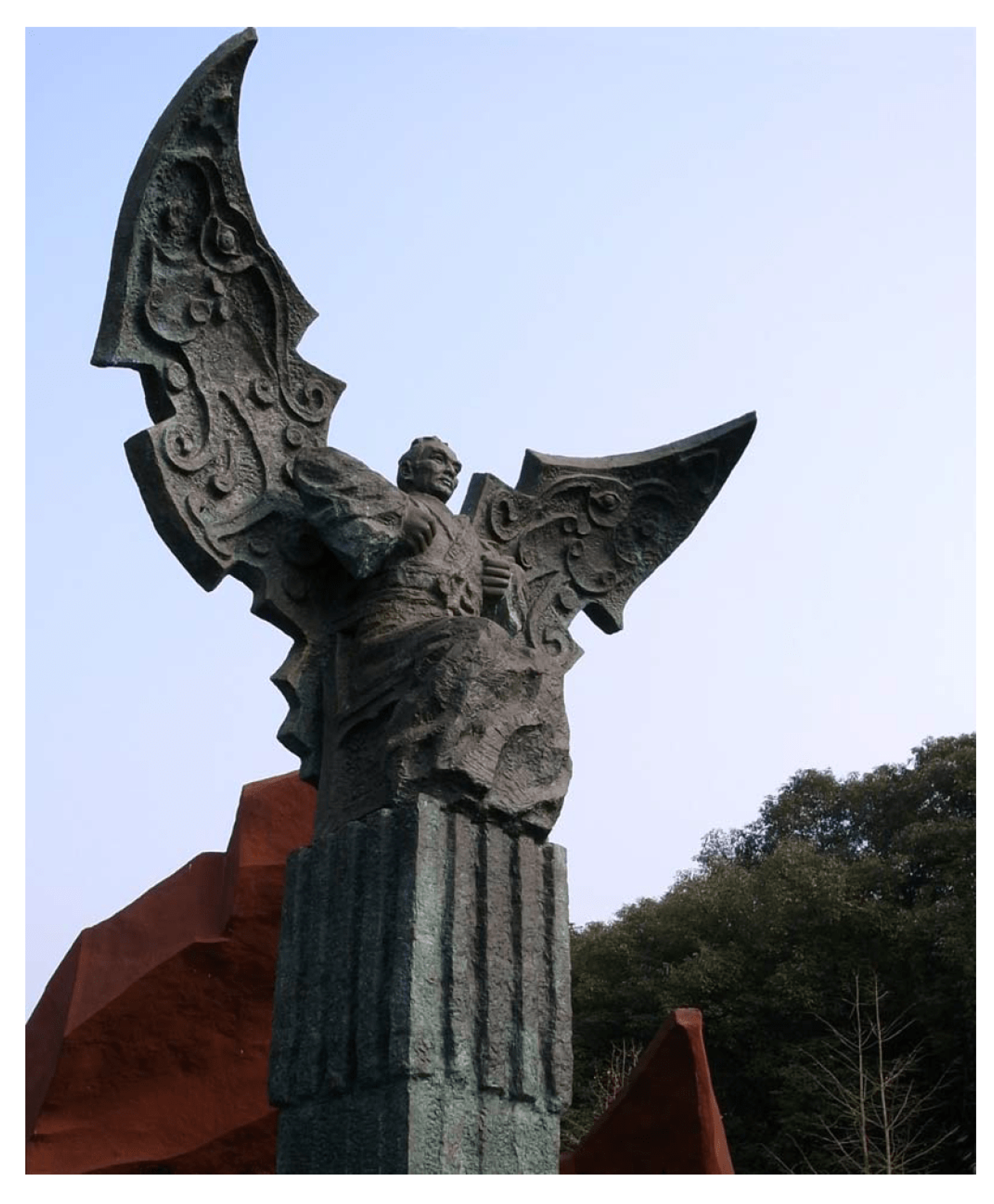
“杨利伟！地球人都知道！”台下有人回应到。

“这个……，还真不是！”徐教授故意拉长声音。

“那是谁呢？”很多人齐声问道。

“万户！”徐教授一边说一边打开大屏幕。

“万户？这个还真没听说过。”马处长感到惊讶。



徐教授扶了扶眼镜，接着说：“既然都没听说过，我就给大家普及下，讲一下万户飞天的故事。”

大家一听到要讲故事，都来了精神。

“明朝初期，有一位木匠出身的官吏万户，做出了一份详尽的计划，他认为按照他的设想，一定能在一个时间段内飞到月亮上去。在这个理想主义者的思维世界里，月亮上是没有人心险恶的……”徐教授讲道。

“因为月亮留下太多美好传说了！”台下有人小声说。

徐教授脸上微露沉重的表情，大家赶紧静下来，徐教授接着说：“他先点燃‘鸟尾’引线，一瞬间，火箭尾部喷火、‘飞鸟’离开山头向前冲去。接着万户的两只脚下也喷出火焰，‘飞鸟’随即又冲向半空，栽了下去。万户虽然失败了，但是他对飞天的探索确实是第一人，万户被认为是人类的航天鼻祖。”

“万户那时候就开始研究火箭了，真是厉害！”马处长甚感诧异。

徐教授说：“人们对挣脱地球引力束缚的欲望一直很强烈。随着科学技术的发展，火箭技术也变得越来越成熟了，但是同时由于各种故障，很多人都为飞天事业献出了宝贵的生命。”

“飞天是全人类的梦想！挫折再多也不能停止前进的步伐！”李部长情绪激昂地说。

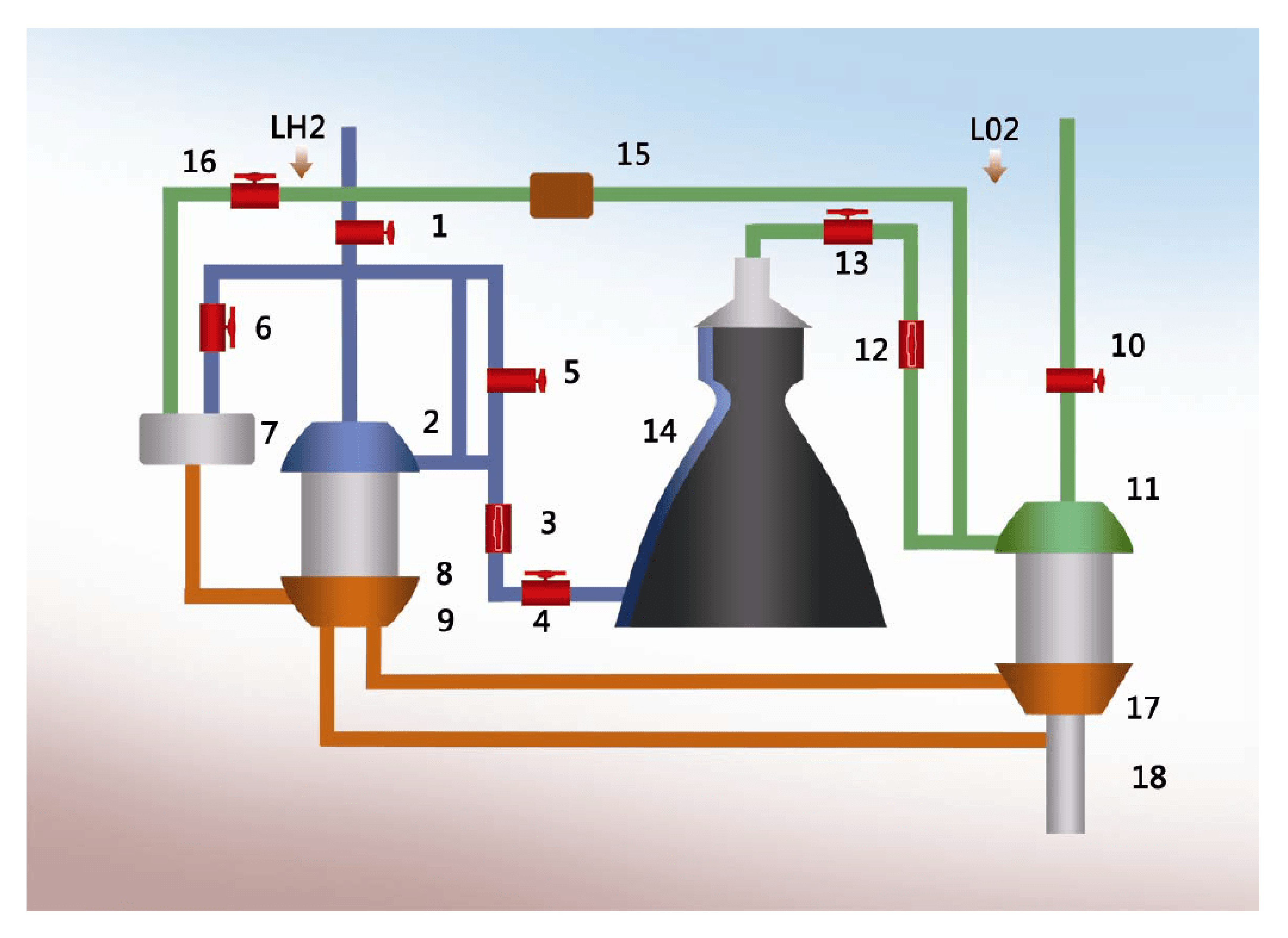
徐教授说：“对，不能停息。但我们不能忘记人类为此而付出的沉重代价：从1959年8月21日美国发射水星号航天器到2006年底，在美国及前苏联/俄罗斯进行的249次载人航天发射飞行中，共出现故障166起，其中最严重的5次载人航天事故包括：1967年1月阿波罗4A号、1967年4月联盟1号、1971年6月联盟11号、1986年1月挑战者号、2002年哥伦比亚号等事故。航天发射事故造成了重大的人员伤亡和经济损失。美国的统计数据显示，动力系统故障占航天器系统总故障的60%以上。”

李部长片刻就明白了徐教授的语义，附和说：“防微杜渐，防患未然。”

“为了让大家更好地了解火箭发动机故障诊断方法，我们还得理解一下火箭原理，这里以当前应用最为广泛的液体火箭为例来说明，请大家看大屏幕。”徐教授调出了一张幻灯片。

“怎么看着像在做化学实验？！”李部长直爽地说。

徐教授详细地解说道：“火箭的主要原理可以用下面的部件描述。1—氢泵前阀；2—氢泵；3—氢主文氏管；4—氢主阀；5—氢副系统旁通阀；6—氢副系统控制阀；7—燃气发生器；8—氢泵涡轮；9—氢氦换热器；10—氧泵前阀；11—氧泵；12—氧主文氏管；13—氧主阀；14—推力室；15—氧稳压器；16—氧副系统控制阀；17—氧泵涡轮；18—排气管。”



“看着挺简单的，咱们可以按图造火箭了！”李部长幽默地说。

徐教授解释道：“我建议大家还是打消造火箭的想法吧，要不你就成为下一个‘万户’了。”

“如何才能提前检测出火箭发动机故障，从而避免事故的发生呢？”电力公司马处长问。

徐教授答道：“液体火箭发动机是一个极其复杂的高能量释放装置，其故障的发生和发展具有极端的快速性和极大的破坏性，其故障的表现也呈现复杂性。这种复杂性体现在环境干扰的多样性、故障特征的多样性、故障的多样性以及内部因素的耦合表现出的很强的非线性，这给液体火箭发动机的故障检测和诊断带来了极大困难。”

“那么当前一般采用哪些方法来进行故障检测与诊断呢？”马处长也表现出了强烈的兴趣。

“常见的故障检测与诊断方法主要有：门限检测方法、信号处理方法、专家系统方法、故障诊断树方法、模式识别方法、模糊数学诊断方法、人工神经网络诊断方法和信息融合的方法等。”

“有了这些方法，可为什么火箭发动机故障还是时有发生呢？”李部长问。

“俗话说，‘金无足赤，人无完人’。基于门限检测的诊断方法由于随机干扰以及各种瞬态过渡过程的存在，使得该方法在检测故障的及时性和准确性方面存在一定困难，且门限值通常难以选取。”

“哦，原来是这样！”李部长回应道。

徐教授看到大家没什么疑惑了，接着讲：“基于数学模型的诊断方法对模型过于依赖，对于参数摄动、噪声干扰等都极其敏感，从而诊断结果的可靠性不能严格保证；基于人工智能的诊断方法需要足够的典型故障样本和先验知识，而现实中发动机故障样本很少，因而这些理论上很优秀的方法难以得到广泛的应用。”

李部长有点穷追不舍的意思，问：“那今天我们用数据挖掘的什么技术来进行故障诊断呢？”

“我们用大家比较熟悉的支撑向量机为例来讲解数据挖掘在火箭发动机故障诊断中的应用。”

“为什么采用支撑向量机呢？”一学员问。

徐教授回答说：“支撑向量机是在统计学习理论的基础上发展起来的一种先进的机器学习方法。它通过最小化经验风险、最小化置信区间的上界以及核函数方法，有效解决了小样本、高维数和非线性以及因样本数较少而带来的‘过学习’问题，克服了神经网络学习方法中网络结构难以确定和存在局部极小值点等缺点，从而具有良好的泛化能力和较强的抗干扰能力。”

“徐老师，支撑向量机在前面的课程里已经讲过了，你就直接给讲应用吧。”李部长提议说。

徐教授也有同样的想法，说道：“好吧。首先要选取合适的故障变量。大量的实际数据中无意义的变量会严重影响数据挖掘算法的执行效率，即引起维数灾难。所以，属性选择是十分必要的。通过深思熟虑，我们从 80 多个测量参数中选取了 22 个变量，具体名称见大屏幕所示。”

参数名称	符号	参数名称	符号
氢泵转速	NWR	氢涡轮入口压力	POWR
氧泵转速	NWY	氧涡轮入口压力	POWY
氢泵流量	GR	氢涡轮出口压力	PEWR
氧泵流量	GY	燃烧室压力	PK
氧泵入口压力	POY	燃烧室氧喷前压力	PY
氧泵出口压力	PEY	燃气发生器压力	PF
氧泵出口温度	TEY	发生器氧喷前压力	PFY
氢泵入口压力	POR	发生器氢喷前压力	PFR
氢泵出口压力	PER	氧泵前活门入口温度	TOHY3
氢泵出口温度	TER	氢泵前活门入口温度	TOHR1
冷却套出口压力	PEL	氧涡轮氢隔离腔压力	Pg

“其次，对采集的这 22 个变量的数据需要认真地进行数据预处理。”徐教授补充道。

马处长问：“数据预处理时，如何对待虚假数据呢？”

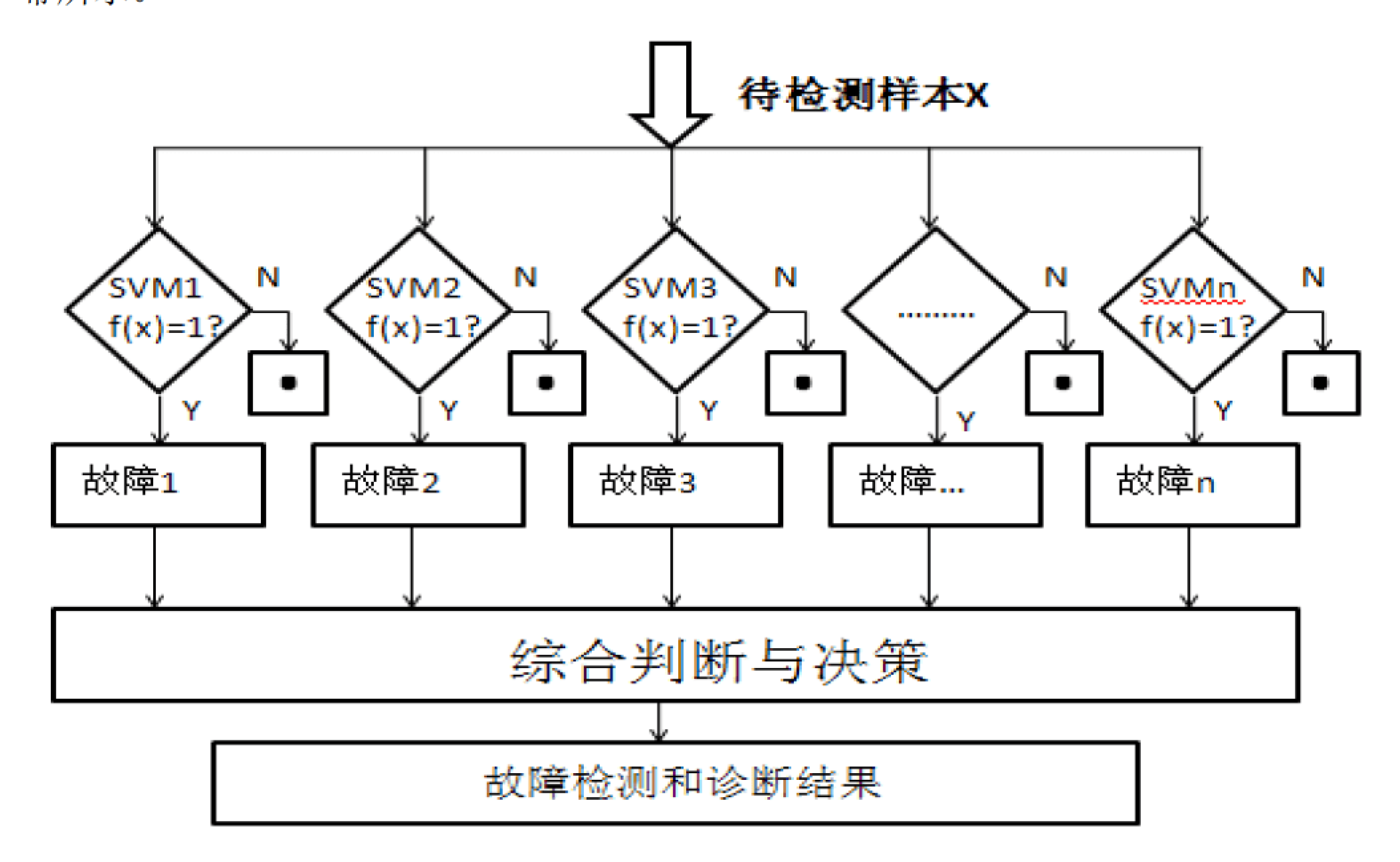
“测量传感器在极端的物理条件下有可能被损坏，使测量的数据无效，因此有时需要剔除严重错误的数据。对于剧烈变化的数据，需分析是传感器测量误差还是真实发生了故障，若是传感器故障，也就是我们所说的虚假数据，则需要对这些数据进行修正。一般采用滑动平均的方法，即选取该点附近一段数据的平均值作为该点的值。”徐教授解释说。

“这个很好理解。”台下有人回应道。

徐教授继续说：“另外，在训练集选取上，要保证数据覆盖范围的全面性。对于正常或发生了同种类型故障的发动机试车数据，不同批次试车，参数值可能都不相同，甚至差别较大，因此在训练集的选取上要尽可能地将不同范围的数据以及反映不同故障类型的数据样本都包含进去。”

“徐老师，支撑向量机是根据二分类问题建立的模型，而火箭发动机故障是多分类问题，如何应用支撑向量机处理多分类问题？”李部长问。

徐教授回答道：“若有 n 种故障类型，则应建立 n 个不同的两类分类器，如大屏幕所示。”



“哦，如此复杂，需要训练这么多的二分类器。”有人说道。

“其实真正应用时，n 一般较小。在本次试车数据挖掘中，共有 3 种类型的稳态段故障模式，氧副文氏管出现多余物、氢涡轮破坏、氢泵次同步振动，加上正常模式共 4 个类别。”徐教授介绍说。

“那就要形成 4 个数据集，训练 4 个 SVM 分类器，是吧？”李部长说道。

“是的，我们在每个数据集中，把属于该故障的样本标号设为-1，其余样本的标号设为 1，然后把每个数据集分为训练数据集和测试数据集，用训练数据集进行训练得到预测模型，然后用测试集对所得到的模型进行预测能力检验。”徐教授说。

“检验效果如何呢？”李部长问。

“训练完成后，就可对测试数据集进行测试了。四种类别的检测正确率都在 92%以上。”徐教授说。

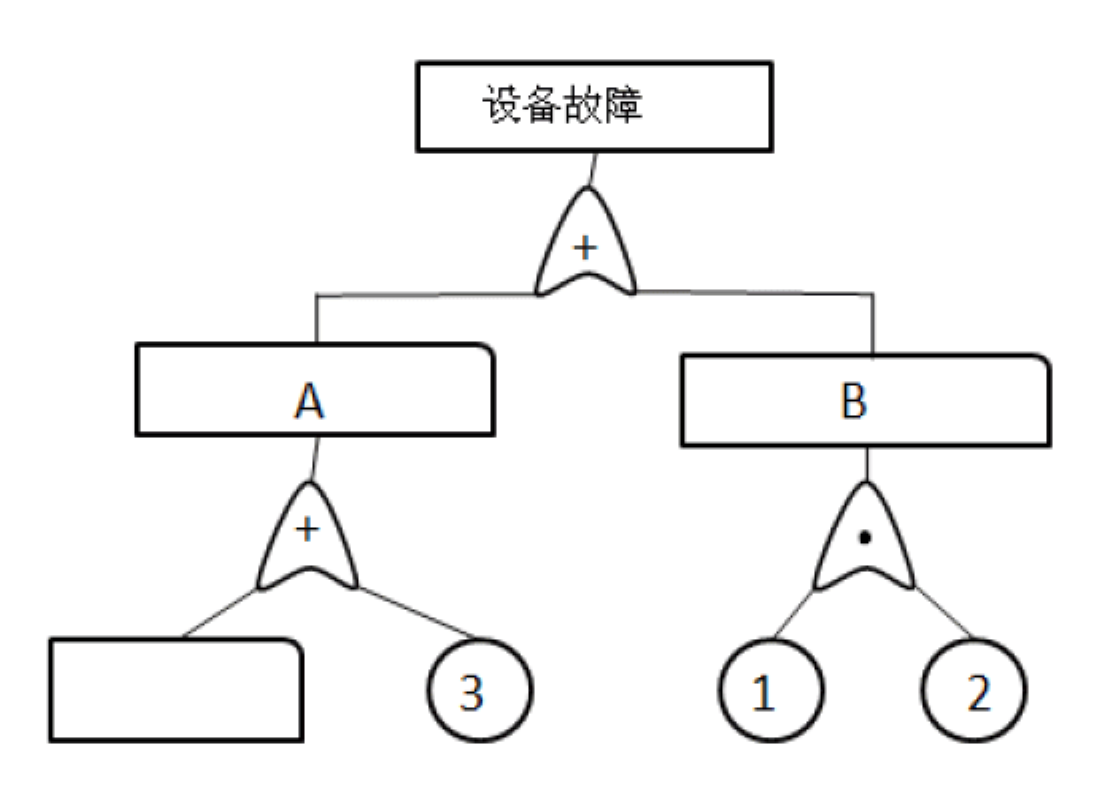
“效果挺不错嘛！”马处长说。

徐教授合上电脑，微笑着说：“这是个试验，只能起到‘抛砖引玉’的作用，还需要进一步研究才有可能在真正实施中应用。今天的课程就上到这里，下次课见。”

7.2　机械设备故障诊断

这节课一开始，徐教授先给大家讲了个笑话：“一个飞机由于机械故障延误了，过了一会又可以起飞了。旅客问为什么？乘务员说没事，就是换了一个敢开的机长。在实际工作中，我们可真不敢这么轻率地对待机械故障，应该响应胡主席提倡的一切以人为本。”

台下另一个学员附和道：“是啊，任何时候，安全都是工作中的重中之重。”



徐教授看着台下的学员，亲切地说道："本节课的主要内容为机械故障诊断。近几年一种故障树分析法 FTA（Fault Tree Analysis）逐渐在实际应用中流行起来。在座的诸位中，谁先给我们说说对 FTA 方法的认识？"

鼓风动力的王总率先说道："故障树分析法是以设备最不希望发生的事件作为分析目标，找出系统内因为环境变化、人为失误等因素导致的部件与部件故障之间的逻辑联系，用倒立树状逻辑因果关系图形表示出来。"

南航的陆经理以前大学念书的时候主修的就是机械设计，说起 FTA 方法毫不费力："故障树是一种从系统到部件，再到零件，按"下降形"分析的方法。它从系统开始，通过由逻辑符号绘制出的一个逐渐展开成的树状分枝图，来分析故障事件发生的概率。同时也可以用来分析零件、部件或子系统故障对系统故障的影响。"

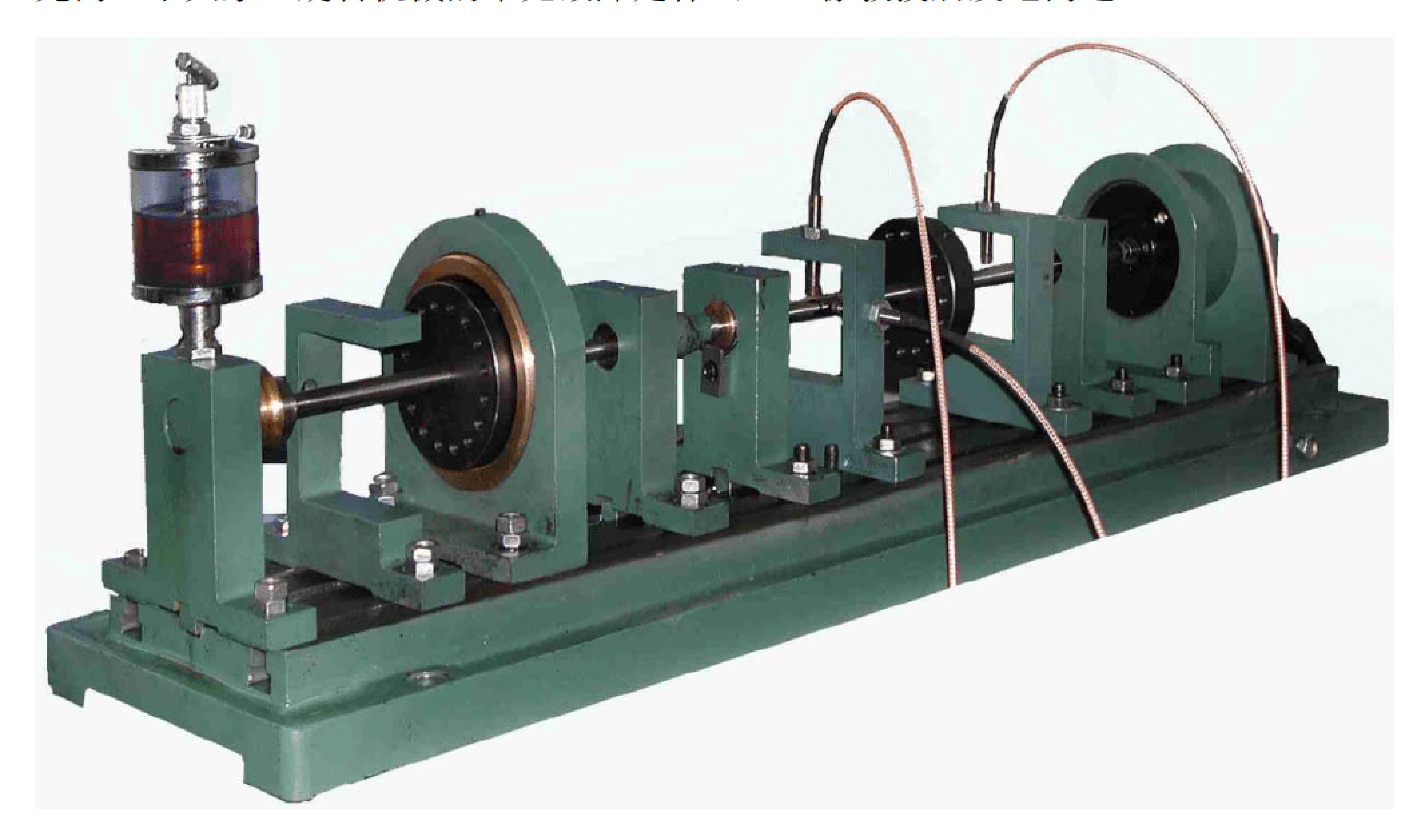
听完陆经理和王总的回答，徐教授十分满意。为了帮助大家更深地理解 FTA 方法，徐教授举了个图例进行说明："它首先选定某一设备故障事件作为顶事件，画在故障树的顶端。再将导致该故障发生的直接原因（各部件故障）并列为第二阶，图上用'或门'表示设备的故障是由部件 A 或者部件 B 故障所引起的；接下来，将导致第二阶的各故障事件发生的原因分别并列为第三阶，如连接部件 B 故障和元件 1 故障、元件 2 故障的是一个'与门'，表明 B 故障是在元件 1、元件 2 同时失效时发生的。对各基本事件赋予先验概率，以表征发生可能性的大小，这样就可以应用故障树模型进行可靠性分析以及诊断决策。"

税务的姚局长听完，笑着问："徐教授，故障树诊断原理我听明白了。FTA 对系统故障不但可以做定性的而且还可以做定量的分析；不仅可以分析由单一构件所引起的系统故障，而且也可以分析多个构件、不同模式故障而产生的系统故障情况。但这个方法有些什么缺点呢？"

徐教授听完后，概括回答道："因为故障树分析法使用的是一个逻辑图，因此，不论是设计人员或是使用和维修人员都容易理解。但是故障树分析法也存在一些缺点，如在构造故障树时要运用逻辑运算，在其未被一般分析人员充分掌握的情况下，很有可能把重大影响系统故障的事件漏掉；同时，由于每个分析人员的研究范围各有不同，其所得结论的可信性也就有所差异。本节课，我们就通过数据挖掘的手段来解决设备故障诊断这个难题。"

"还是按照老规矩，给我们结合实际例子讲讲吧。"台下有人建议说。

"好的。在座的当中，生产企业的同志不少，我们今天就以旋转机为例进行讨论。先问一下大家，旋转机械的常见故障是什么？"徐教授启发地问道。



"转子不平衡。"李部长立刻喊道。

"那引起转子不平衡的原因又是什么呢？"徐教授接着问。

"装配不规范或机械磨损所致。"李部长又答道。

徐教授接着道："不错，当转子不平衡时，振动的时域波形为正弦波，谐波能量主要集中于基频，且工作转速一定时相位稳定。此外，转子的轴心轨迹为椭圆，这时候，振动的强烈程度对工作转速的变化非常敏感。还有什么引起转子不平衡的原因？"

马处长说道："油膜涡动。比如轴承发生油膜时，尽管振幅较小，对轴承的润滑和工作影响不大，但它所产生的附加动力载荷容易使机器零部件发生松动和疲劳失效等故障。"

R 钢铁公司的何总说道："马处长说起油膜，提醒了我。油膜振荡，是轴颈带动润滑油高速流动时，高速油流反过来激励轴颈，使其发生强烈振动的一种自激振动现象。"

S 钢铁公司的赵总也补充道："还有转子支承系统连接松动和转子不对中，也可引起转子不平衡。"

李部长又想起了一种原因，说道："还有喘振，它是透平压缩机特有的现象。喘振较大时常导致转子弯曲、联轴器及齿轮箱损坏等。"

徐教授进一步说："转子不平衡的原因基本上就是这些了，要诊断出这些原因，就必须选取故障特征变量。我们选取故障信号的频率特征、振动特征、敏感参数作为故障识别的标准故障模式，组成故障识别参数集。"

"徐老师，复杂工业系统的设备繁多，系统复杂，经常出现多种故障原因同时作用，应用常规的分类方法难以进行这样的故障诊断吧。"李部长说出了自己的想法。

徐教授高兴地说道："李部长分析地很对。对于这类问题，通常用建立故障变量与故障类型之间的多值关联规则的方法进行故障诊断。在机械设备故障诊断中，关联技术就是寻求设备中各因素间的主要关系，找出影响目标值的重要因素。从而掌握事物的主要特征，促进和引导系统迅速而有效地发展。"

"徐老师，我记得关联规则只能使用离散型数据，而这个故障诊断问题数据是连续的。"李部长分析说。

徐教授说道："对。使用关联规则算法前，需要对连续型数据离散化。我们首先为每一连续型属性划分为几个区间段，然后把该属性的值影射到各个区间。比如，在识别参数集中，各段频率都是经过离散化，转换为离散值，其他振动特征与敏感参数（如相位特征、轴心轨迹、转速、油温等）的表征为：稳定、较稳定和不稳定，或者规则的、较杂乱、杂乱和杂乱并扩散，或者不变、不明显、有变化、明显和很明显。"

随着徐教授的详细解释，大家明白了数值预处理过程。

徐教授接着描述道："接着计算所有属性经过划分后的支持度，如果出现比支持度阈值小的情况，则考虑重新划分或合并相临区间。找出比最小支持度大的所有项集，得到频繁项集，最后就可以由频繁项集提取故障诊断的关联规则。"

有人又问道："发现关联规则后，怎么从数据来判断故障类型呢？"

徐教授回答道："如果新来故障的数据满足不平衡故障规则所提供的条件，我们就可以判定该条数据为不平衡故障。例如，如果新来故障的数据满足松动的故障规则所提供的条件，我们就可以判定该条数据为松动故障。同样，这种方法适用于其他故障数据。

7.3　核动力设备故障诊断

徐教授提着公文包迈进教室，打开笔记本电脑，将 PPT 停留在第一页幻灯片上。

李部长看到大屏幕上有一张核电站的图像，旁边画了一个大蚂蚁，不知徐教授的意图。

大家陆陆续续到了，徐教授清了清嗓子，说："今天将讲解数据挖掘技术在核动力设备故障诊断中的应用。"

听到"核动力设备"这几个字，大家顿时安静不住了。因为前不久的日本福岛核泄漏事故给大家留下了很深的印象，特别是"盐荒"事件。

徐教授接着说："其实严重的核泄漏事件不只是日本福岛核事件，比如，1979年3月发生的美国三里岛核电厂2号机组事故就是由于操作人员未能识别出稳压器卸压阀未关闭，并执行了错误的动作而导致；1986年4月苏联切尔诺贝利4号机组事故主要原因是对运行规则的粗暴违反，这两起事故都造成了反应堆烧毁，放射性物质外泄。另外2004年8月9日日本关西电力公司位于福井县美滨核电站3号反应堆发生涡轮机房内蒸汽泄漏事故，虽然没有放射性物质泄漏，但造成了4人死亡，7人受伤的后果。"

"现在大家让日本核泄漏事件搞得是谈'核'色变了都！"李部长略带气愤的语气说。

徐教授说："我们不能怕了'核'就不利用它了，相比较起来核能源是非常安全和绿色的能源。就比如，我们不能怕噎着就不吃饭了不是，哈哈！"

"徐老师，您先给介绍下核动力装置到底是怎么样的一个装置吧。"马处长不愧是个急性子。

徐教授解释说："核动力装置是一个技术密集、结构复杂、造价昂贵的复杂系统。在纵向，核动力装置可按层次分解为多种不同类型的系统或设备；在横向，诸多设备之间通过功能接口与物理接口关系、控制关系相互保障和制约，构成一个有机整体；由于核安全的限制，还存在纵深防御、多层屏障。"

"确实比较复杂！"台下有人说。

"又由于其依靠核反应堆来提供动力来源，因此核动力装置的特殊性不仅表现在其复杂性上还表现在其发生故障后可能的潜在放射性危险上。"徐教授说。

"日本福岛核泄漏是个大教训啊！至少我不想再经历一次'盐荒'了！"李部长意味深长地说。

徐教授幽默地把"盐荒"放在核动力安全之前，说道："为了不再发生'盐荒'，同时适应核动力装置安全性和可靠性的更高要求。各种先进技术和设备得到广泛使用，但是也使得核动力装置越来越复杂，这也给操纵人员带来了很大的困难。"

马处长点头表示赞同说：“有道理！”

徐教授接着说：“核动力装置主控室内的报警量就有 2000 多个，一旦发生故障，可能会出现多个参量同时报警，这些报警信息虽然可以帮助操作人员进行故障辨识，但同时也给操作人员带来很大的压力，从而影响其做出正确的决策，核发展史上的几次重大事故也证明了这一点。仅凭操作人员的技能和经验是不能很好的对核动力装置进行控制的。”

“是的，‘经验主义害死人’！”李部长说。

徐教授说：“越来越多的人认识到经验是靠不住的，所以随着核技术的不断发展和应用以及人们对核安全的高要求，如何保证核动力装置的安全运行受到了核能界的高度重视，研究人员在尽量提高核动力装置自身的固有可靠性的同时开始注重开发核动力装置故障诊断系统。”

“核动力装置故障诊断系统？”有人有点怀疑。

徐教授说：“是的，核动力装置故障诊断系统是一种操作人员的支持系统，其目的是使故障诊断更容易、更准确，降低事故时的人为失误并减轻事故给操作人员带来的压力，提高系统的可靠性和有效性。”

“这样的系统太亟需了！”马处长联系到自己行业的大型设备言道。

“那核动力故障诊断系统的目标是怎样的呢？”李部长问。

徐教授回应说：“核动力故障诊断系统能够对装置的主要运行参数进行监控，或发现核动力装置可能的运行故障，发出相应的报警信息或给出故障的部位、故障的原因，使操作人员能够及时地发出命令，防止出现运行故障，使核动力装置具有较好的运行状态，从而达到改善核动力装置运行性能、保证其安全性、提高核动力装置的易操纵性的目标。”

“要求这么高，实施确实有难度！”台下有人道。

“不怕，我们可以用这个！”徐教授翻动幻灯片，屏幕上出现一个巨大的蚂蚁。



“蚂蚁！”大家异口同声地喊到，并感到很纳闷。

“是的，大家可别小看蚂蚁哦！”徐教授看到坐在后排的同学精神不济，就打算给大家提提神。

徐教授接着说：“蚂蚁很强大哦！我先给大家讲个笑话！”听到徐教授讲笑话了，后排的同学顿时来了兴致。

徐教授托了托眼镜，说：“一只蚂蚁在路上看见一头大象，蚂蚁钻进土里，只有一只腿露在外面。小兔子看见不解地问：‘为什么把腿露在外面？’蚂蚁说：‘嘘！别出声，老子绊他龟儿子一跤！’第二天，兔子看见整窝的蚂蚁排着队急匆匆赶路，问何故，蚂蚁答：‘昨天有头大象被我们一个兄弟绊倒，摔成重伤，我们给那厮献血。’没多久，兔子见大批蚂蚁又回来了，就问怎么回事，一只蚂蚁说：‘哦，只有一只蚂蚁跟那大象的血型一致，留他一个在那抽血呢。’”

听完关于蚂蚁的这个笑话，大家被逗得前仰后合。

“蚂蚁强大吧？！至少让大家不瞌睡了！”徐教授幽默地说。

看到后排瞌睡的学员又有了精神，徐教授接着说：“今天我们的主角就是蚂蚁，我们要用蚁群优化算法进行故障诊断！”

“一个蚂蚁都够厉害，如果是蚁群就更厉害了，那不要绊倒一群大象？！”李部长开玩笑说。

“我们在蚁群算法中提出了人工蚁的概念。人工蚁有着双重特性，一方面，它们是真实蚂蚁行为特征的一种抽象，通过对真实蚂蚁行为的观察，将蚁群觅食行为中最关键的部分赋予了人工蚁；另一方面，由于所提出的人工蚁是为了解决一些工程实际中的优化问题，因此为了能使蚁群算法更有效，人工蚁具备了一些真实蚂蚁所不具备的本领。”徐教授解释说。

“俗话说思想照多远，我们就能走多远，那蚁群算法是怎么个思想呢？”李部长幽默地说。

“蚁群算法的基本思想可概括为，在蚁群优化算法中，一个有限规模的人工蚁群体可以相互协作地搜索用于解决优化问题的最优解。每只蚂蚁根据问题所给出的准则，从被选的初始状态出发建立一个可行解，或是解的一个组成部分。每只蚂蚁都能够找出一个解，但很可能是较差解。蚁群中的个体同时建立了很多不同的解决方案，找出高质量的解是群体中所有个体之间全局相互协作的结果。”

“具体该怎么理解呢？”

徐教授解释说：“蚂蚁在觅食过程时，是以信息素作为媒介而间接进行信息交流，当蚂蚁从食物源走到蚁穴，或者从蚁穴走到食物源时，都会在经过的路径上释放信息素，从而形成了一条含有信息素的路径，蚂蚁可以感觉出路径上信息素浓度的大小，并且以较高的概率选择信息素浓度较高的路径。”



“哦，有点深奥！”李部长挠挠头，表示不太理解。

徐教授接着解释说：“蚂蚁在路径上前进时会根据前边走过的蚂蚁所留下的分泌物选择其要走的路径。其选择一条路径的概率与该路径上分泌物的强度成正比。因此，由大量蚂蚁组成的群体的集体行为实际上构成一种学习信息的正反馈现象。”

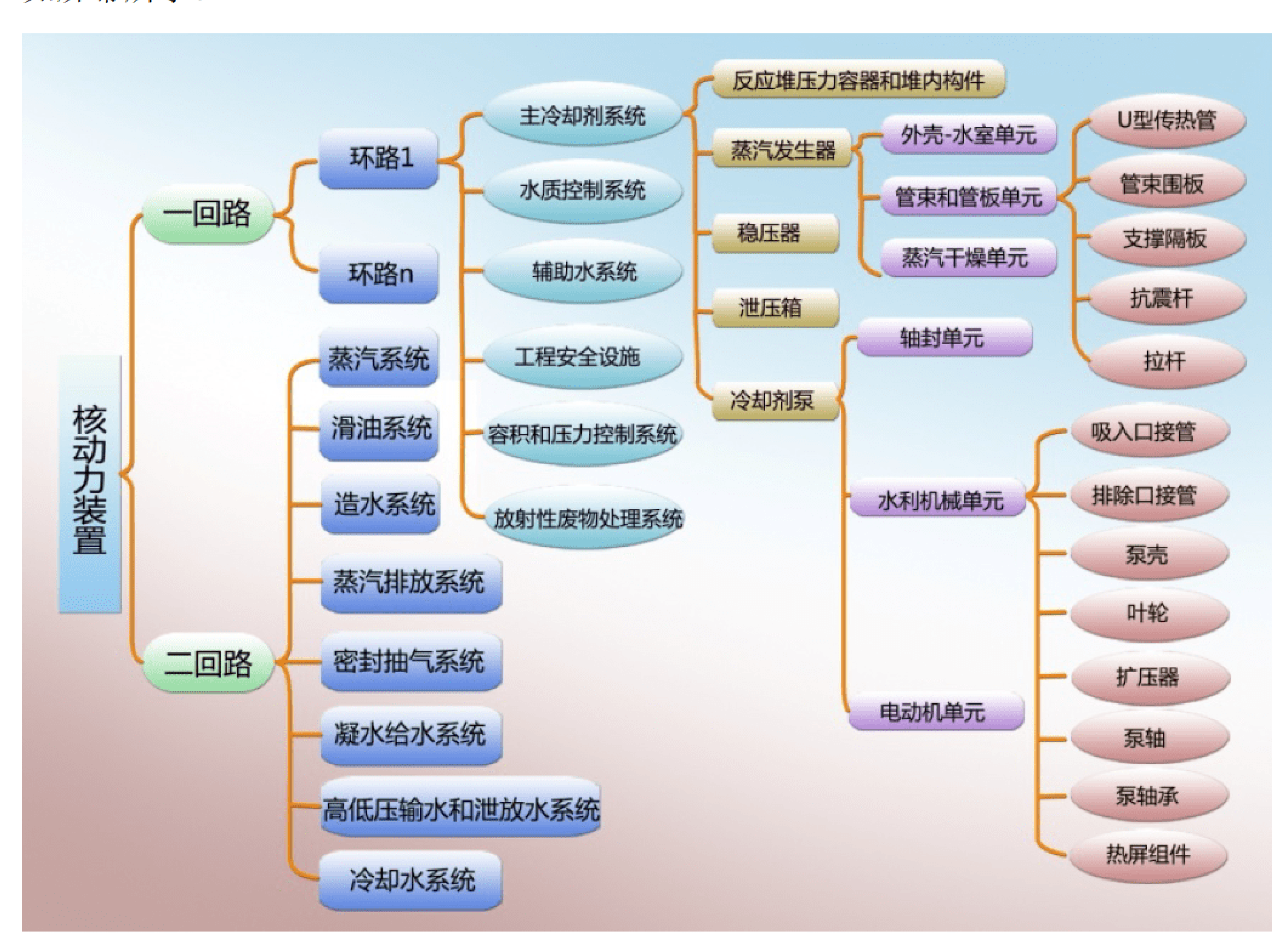
“什么现象？”马处长探着脑袋问。

“某一条路径走过的蚂蚁越多，后面的蚂蚁选择该路径的可能性就越大。蚂蚁的个体间最终通过这种信息的交流寻求通向食物的最短路径。”徐教授说。

“不错，蚂蚁虽小，给人类的启发作用确实很大啊！”李部长说。

徐教授："这里给大家讲的是最初的蚁群算法，该算法还存在很多不足，有很多学者对蚁群算法进行了改进，例如蚁群系统算法等，这些我们就不在课堂上详细讲解了。"

徐教授看到大家对蚁群算法的基本原理大概接受了，接着说："对于在核动力设备故障诊断中的应用，我们首先将核动力装置分成各级，对不同级别进行故障定位。由于算法及分析的复杂性限制，今天以一回路的主冷却剂系统为研究对象，将其划分到设备级进行故障定位研究，各设备简化为节点形式，简化后的一回路主冷却剂系统如屏幕所示。"



"虽然简化了还是蛮复杂的哦！"李部长趴在马处长耳畔低声说。

徐教授接着解释道："由于核动力装置发生故障后，对于同一个参量，可能不同设备发生故障时，均会变为异常，例如稳压器水位，当稳压器自身发生故障时会偏离正常值，当蒸发器或主冷却剂管道破裂等故障发生时，它也会偏离正常值，因此对于不同设备的状态描述可能涉及相同的参量。"

"这就导致故障定位比较麻烦了。"李部长认为。

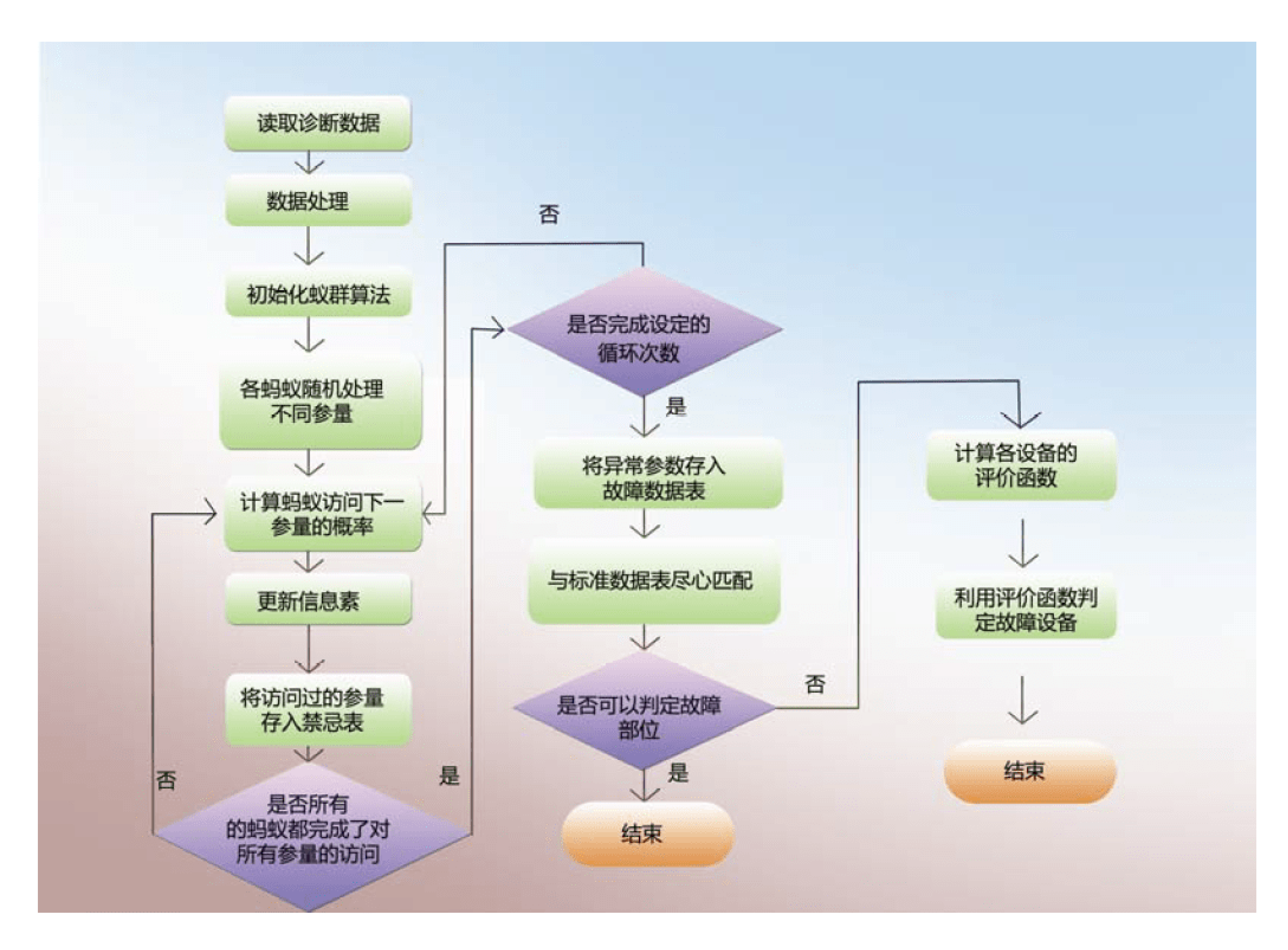
徐教授翻动幻灯片，说："我们选择与所要研究的对象相关的故障。屏幕显示的为 100%功率运行发生故障时状态参量变化的最大限值"。

参量	含义	单位	参量	含义	单位
1	堆芯流量	%	9	蒸汽流量	Kg/s
2	热管段温度	℃	10	蒸发器破裂流量	Kg/s
3	冷管段温度	℃	11	一回路流量	Kg/s
4	堆芯流量	Kg/s	12	稳压器压力	Mpa
5	蒸发器水位(宽量程)	%	13	热功率	%
6	蒸发器水位(窄量程)	%	14	核功率	%
7	蒸发器压力	Mpa	15	过冷裕度	℃
8	给水流量	Kg/s	16	稳压器水位	%
			17	平均温度	℃

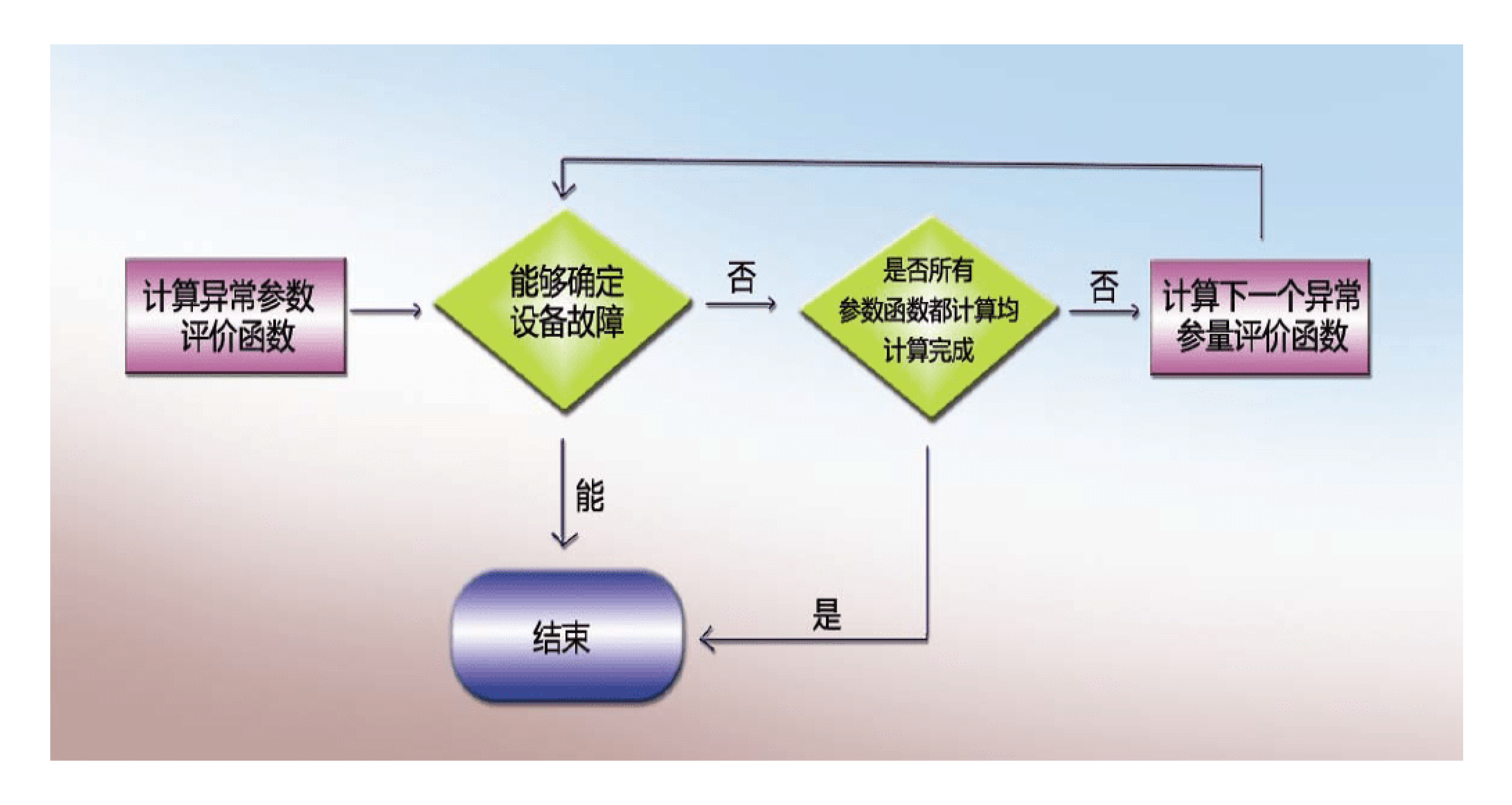
设备	故障/参量	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
反应堆	破口事故	03	06	08	217	02	02	01	40	27	35	04	02	03	03	03	13	05
蒸汽发生器	传热管破裂	05	15	13	639	04	03	01	29	60	28	03	02	04	03	02	09	07

"那么是如何实现故障定位的？"台下有人问。

"采用蚁群算法实现故障定位的基本流程如屏幕显示。"徐教授说。



徐教授学着爱问问题的李部长语气说：“如何通过评价函数来进行故障定位的呢？”惹来大家一阵笑声，然后接着说：“我特意做了一个流程图，通过这个图大家可以清楚理解是怎样通过计算评价函数来实现故障定位的。请大家看大屏幕。”

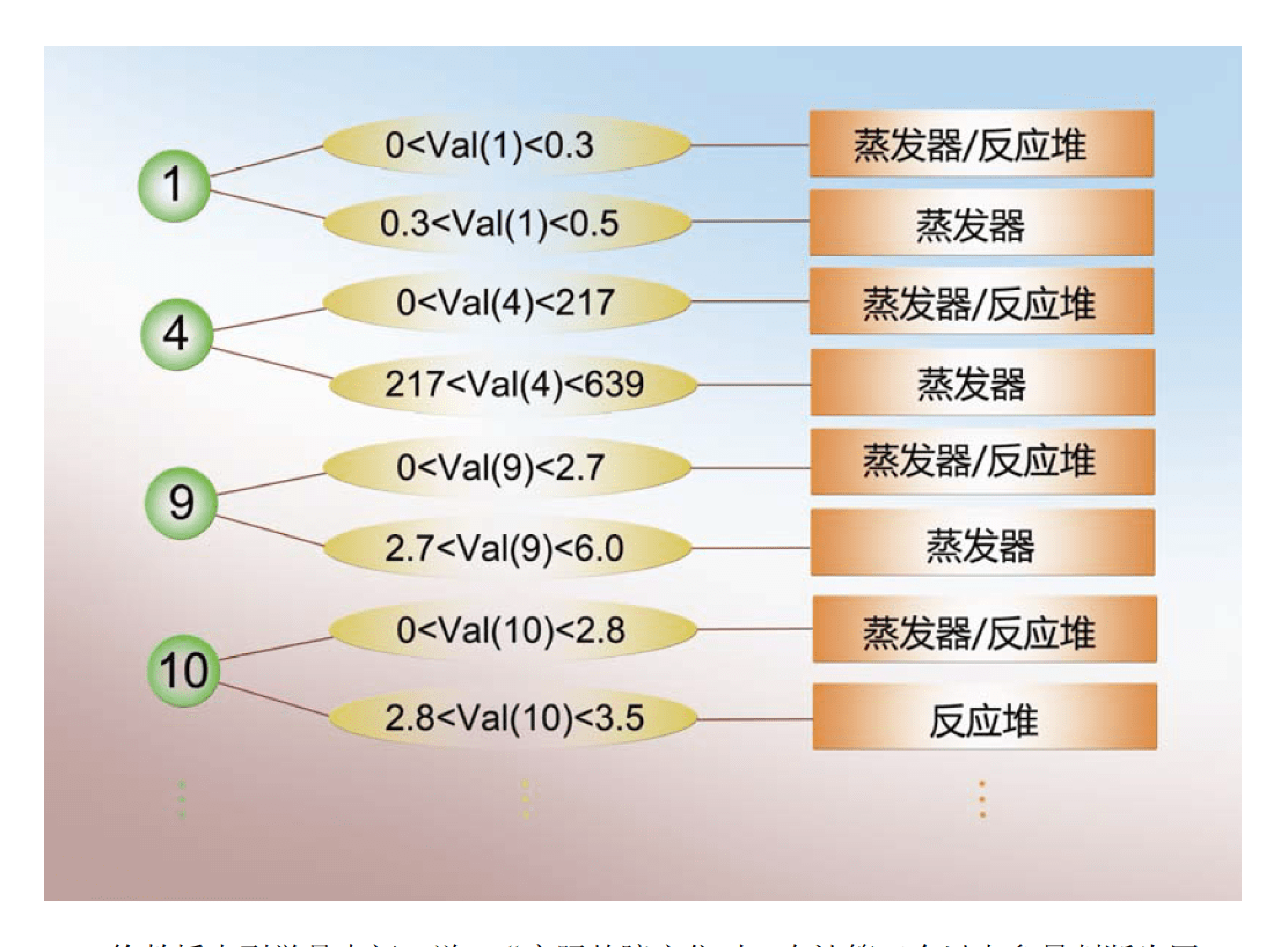


“那么蚁群算法的结果怎么样？”李部长问。

徐教授看到大家好像都有这样的疑惑，就解释说“蚁群算法搜索出的异常参量列表为 1→4→9→10→8→2→3→17→16→5→12→6→13→11→15→14→7。从蚁群算法搜索出的异常，参量列表可以看出偏离正常值较大的参量都排在列表前面，这对于快速判断故障设备是有利的。”

李部长恍然大悟：“原来是这样！”

徐教授接着说：“核动力装置在运行的过程中，设备发生故障时，一般情况下不同设备异常，参量也不相同，因此大多情况下直接利用蚁群算法搜索异常参量与标准故障表进行比对即可。当不同设备发生故障，异常参量也相同时，就利用评价函数一一对各参量进行计算，通过评价函数可确定故障设备。”



徐教授走到学员中间，说：“实际故障定位时，在计算三个以上参量判断为同一设备故障后给出最终判断结果。当所判断故障设备越多时，参量评价函数的分级越多，由于这里只考虑了两个设备，所以对评价函数分级时只分为两级就可以判断故障设备，从而完成故障定位。是不是很简单？”

“徐老师，这样就好，您点到为止就好，让我们明白道理就可以了，再深入了我们接受起来就有困难了。”电力公司的马处长说出了大家的感受。

“好，不过我有信心让大家能够了解数据挖掘在故障诊断中的应用方法。这节课到此结束！”徐教授挥手示意下课。

7.4 船舶动力故障诊断

“有一首歌很火，叫《纤夫的爱》，相信在座的都很熟悉。”徐教授开场道，随后将这首歌曲的前面几句歌词“妹妹你坐船头，哥哥在岸上走，恩恩爱爱纤绳荡悠悠……”写在了黑板上。

有人说道：“这个歌真是太有意思了，看着歌词都没法念出来，念着就想唱。即使不念，也一定会是在心里唱着出来。”



徐教授说道：“这个歌确实有意思，哥哥的纤绳荡悠悠之后，妹妹的船就走动了。由此，引出了我们今天要讲的：船舶动力问题。讲到船舶动力，就不能不提船舶动力装置。”

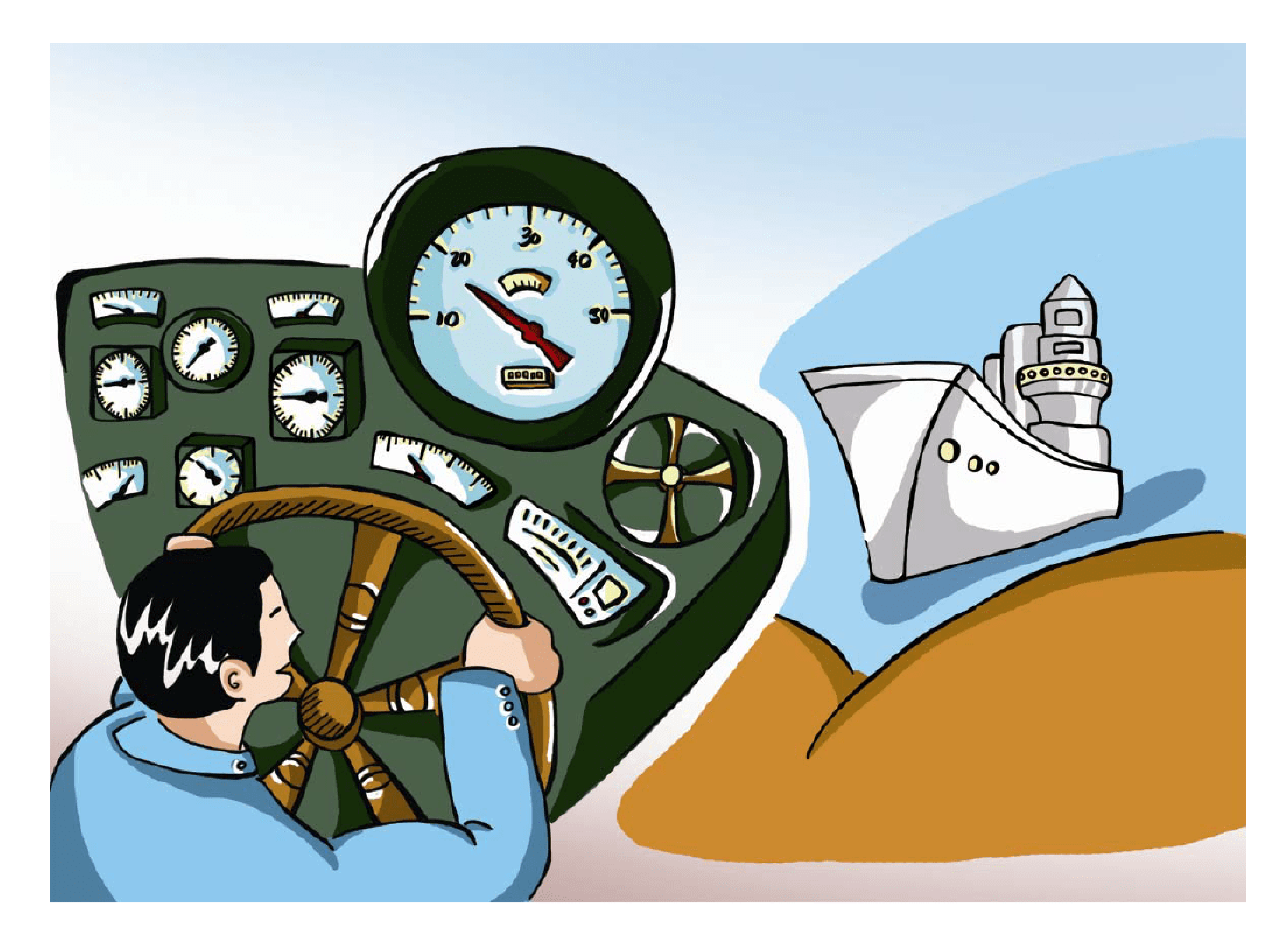
台下一个学员道："徐教授，船舶动力装置很好理解，就是为保证船舶正常营运而设置的动力设备。它是为了保证船舶正常航行提供能量的机械设备，应该说是船舶重要的组成部分。"

徐教授继续说道："是的，一般地船舶动力装置包括主动力装置、辅助动力装置等。主动力装置包括主机、传动设备、轴系、推进器及其附属设备，是全船的心脏。辅助动力装置包括为全船提供电力、照明和其他动力的装置，如发电机组、副锅炉等。主动力装置以主机类型命名，比如蒸汽机类的。"

台下有人问道："世界上有两条船最出名，一个是诺亚方舟，另一个是泰坦尼克号。诺亚方舟我们都知道是神话传说，就不说了。我想问的是那个泰坦尼克号的动力装置是什么类型的？"

说起这个话题，马处长站起来说："泰坦尼克号以煤为燃料产生蒸汽推动蒸汽机工作，船上有 25 台双端锅炉和 4 台单端锅炉，它们的动力来自 159 台煤炭熔炉，它们 24 小时源源不断地为泰坦尼克号提供维持强大动力的蒸汽，动力系统由 3 套主机组成，其中 2 套为 4 汽缸往复式蒸汽机，另外 1 套为蒸汽轮机。"

"除了蒸汽机类的动力装置，还有汽轮机、柴油机、燃气轮机和核动力装置等几类船舶动力类型。"徐教授接过话题说道："船舶动力设备由于结构复杂、工作条件恶劣等原因，发生故障的几率较高。若忽视其状态监测与故障诊断，很可能造成难以想象的重大事故。"



国内某船舶所的程主任感慨地说道：“徐老师，您也知道，由于船舶动力设备是复杂的非线性系统，对于我们这些不懂新技术的‘老古板’，在实际中要准确界定设备故障所在真的是十分困难。我们单位每年因动力设备故障造成的经济损失都非常巨大，还可能造成重大的人员伤亡，因此如何及时发现和排除故障意义十分重大。”

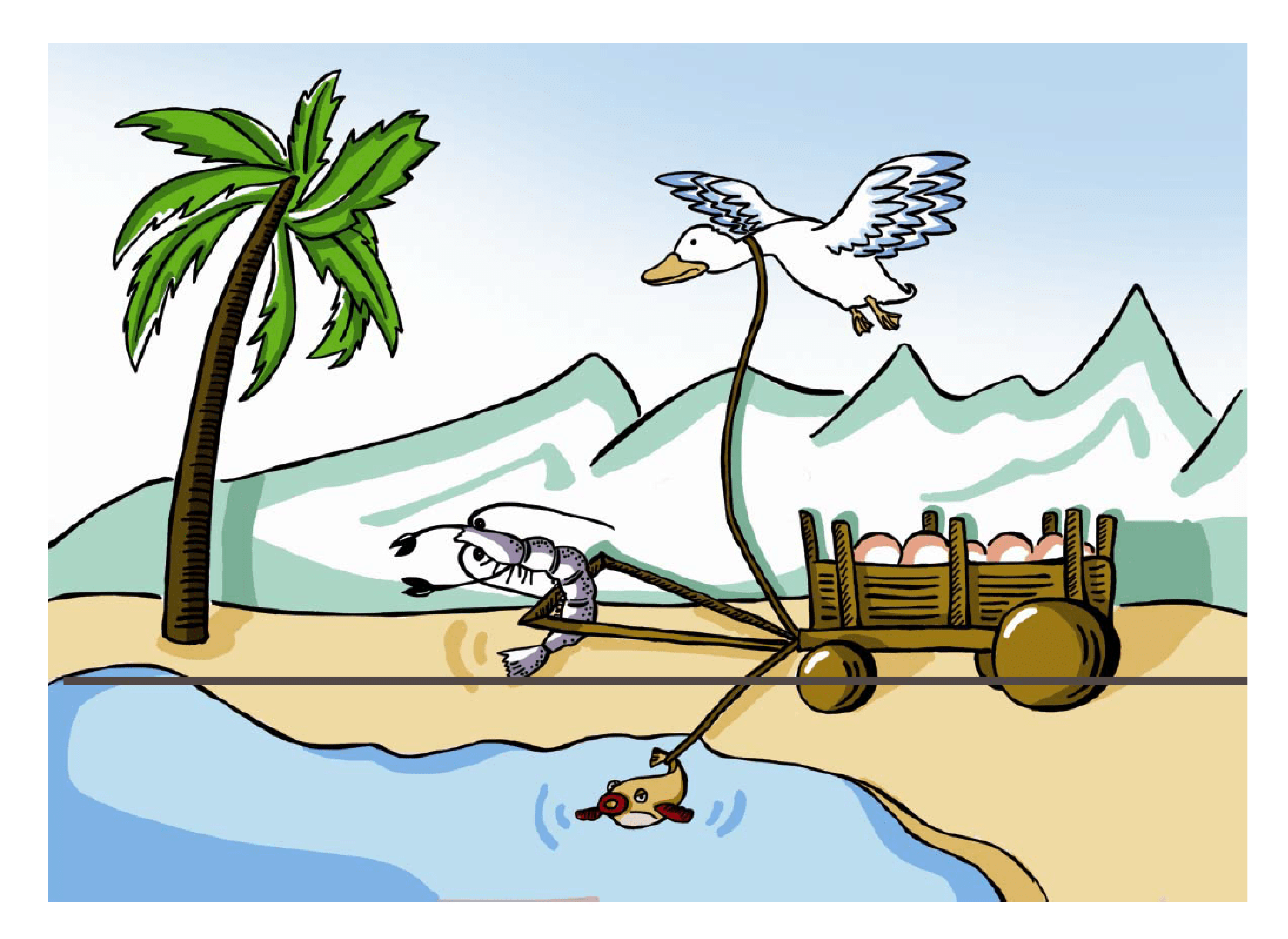
“徐教授，我有个问题咨询下程主任。”得到徐教授手势示意后，马处长问道：“程主任，我们电力行业在实际中的故障诊断一般都是通过计划检修来发现的。虽然目前已经有一定程度的在线监测，但是对收集来的信息利用还是很低的。前面徐教授提过状态检修来改善这个窘况。不知道你们船舶业的状态检修是个什么情况？”

程主任回答道：“说来惭愧，我们目前的维修作业多采用预防性维修和事后维修。预防性维修，你也知道，就是对没有发生故障的设备确定一个强制性的维修计划，使每台设备都有自己固定的维修保养周期，再就是在设备已经出现故障后的事后维修。”

台下的华润万家的万总站起来说道："国内的情况都差不多。开展状态监测维修的基础是用可靠的方法获取到设备的真实状态，为此需要相应的设备检测诊断系统的支持。它是根据设备的日常点检、定期检查、状态监测和诊断提供的信息，经过统计分析处理来判断设备技术状态的好坏，并在故障发生前有计划地进行维修。随着技术发展，我们目前已经有了这个开展状态维修的基础了。"

听完大家的意见，徐教授说道："研究领域，国内多年来热衷于算法的改进，始终未能形成工程性的实用产品。根据近年本领域发表的大量文献看，当前，我国船舶动力设备诊断系统的相关研究与工程应用的实际情况有较大差距。造成该现状的主要原因是缺乏一个整体性的资源平台，无法形成研究的合力。"

徐教授说："没有形成合力这点像《天鹅、大虾、梭鱼拉车》的故事：天鹅、大虾、梭鱼想拖着一辆大车跑，它们都给自己上了套，拼命地拉呀拉呀，大车却一动也不动，车子虽说不算重，可天鹅伸着脖子要往云里钻，大虾弓着腰儿使劲往后靠，梭鱼一心想往水里跳。究竟谁是谁非，我们管不着，只知道，大车至今仍在原处，未动分毫。"



台下一个学员问道："真形象，徐教授，想做出一番贡献还必须得靠合力。那目前我们形成的合力以推'故障诊断数据挖掘'这辆大车的力量有哪些呢？"

徐教授回答说："据故障诊断技术向智能化、综合化、系统化方向发展趋势的特征，工程船舶动力机械研究成果主要体现在将热力参数分析法、油液分析法和振动诊断法等多种诊断方法综合应用；将新的信号分析和处理方法（如神经网络和遗传算法等数据挖掘技术新技术）应用于柴油机信号的分析与处理中，开展工程船舶动力机械智能专家诊断系统的研究。"

"徐教授，还是给大家具体讲一个方法来说明一下吧。"台下有人建议道。

徐教授回答道："对机械设备润滑油进行光谱分析，是故障诊断的一种手段。光谱分析数据是被监测油样中包含的 19 种金属和非金属元素的质量浓度值。聚类分

析方法利用油样所有元素浓度的分布情况，计算油样之间的距离关系，来确定油样状态进而确定故障原因。”

台下有人问道：“这里需要考虑的数据指标有哪些呢？”

程主任抢先回答道：“这个我知道，如油液分析中的光谱分析，获得 Fe、Al、Cu 等 19 个元素的浓度值，这 19 个元素浓度指标就是 19 个基础指标单元。也可以使用 19 个基础指标单元延伸出的变化率，比如 Fe 的变化率、Al 的变化率、Cu 的变化率等。”

王经理也说出自己的看法：“行业某些问题是共通的。所以，我想在实际诊断过程中要用到大量能够数字化量度的状态判断判据或定性规则。如振幅多大是振动过大，相位变化多少属于不稳定或稳定，诸如此类。有很多这类需要界定的判据，这是整个诊断是否正确的一个关键环节。”

徐教授继续说道：“程主任业务果然很精通，王经理也说得很对。经过这些数据准备之后，就需要经历聚类技术的两个步骤。第一步就是数据标准化：常用的方法是平均绝对偏差法。平均绝对偏差法主要包含计算每个属性的平均绝对偏差、标准化的度量值。第二步是计算相异度：数据间的相异度是基于对象间的距离来计算的（常用欧几里德距离）。”

台下学员举手问道：“我经常看见说数据预处理时需要数据标准化，但是不明白原因，为什么要标准化呢？”

旁边的刘经理回答道：“我的理解是：说通俗点就是两字‘平等’：让属性数据处在同一起跑线上，然后再进行分析。在需要聚类的样本由多个属性组成时，不同属性的绝对值变化范围可能因为量纲的关系相差很大，为此需要对属性值进行无量纲化处理。”

徐教授点评道：“刘经理回答得很好。通过聚类方法，对实际运行的柴油机不同时刻采集的 13 个油样光谱油料进行分析。通过油样状态的聚类结果，就可以辅助判断出柴油机的故障原因。”

“哦，我明白了。通过聚类形成的一定的准则及诊断策略将特征提取获得的待检模式与数据库中已有的故障案例进行对比分析，就能识别设备当前所处的状态。但是很多时候数据库这些故障案例累积是很少的，极端情况下可能一个都没有。徐教授，这个问题有什么好解决办法没？”

徐教授说道：“这个问题问得很好。在进行聚类算法有效性验证的时候，比较令人困扰的是：故障诊断的判别数据（也就是测试集）的获取非常困难，要想从单一的运行参数中获得可供诊断的判据数据几乎是不可能的事情。这时候，利用挖掘跨国公司客户服务数据库中的服务数据来提炼诊断判据知识，能突破这种数据匮乏瓶颈。”

“这个数据共享技术实现难度高不高呢？”台下有人说出自己的疑问。

徐教授回答道：“已经证实可以实现，某单位已经利用数据库技术、远程通信技术和模式识别等信息技术来解决判据获取的难题。”

王科长激动地说：“这样大力地改进船舶动力设备故障诊断技术后，提高了系统运行可靠性，最主要的是保证了参与者的生命安全。”

张经理也附和道：“是啊，这样能更准确、更及时地了解设备状态，使零部件的性能得到充分的利用，降低维修费用，从而获得最佳经济效益。”

鼓风动力集团的王总也高兴地说道：“可不是么，好处一点都不可小视。船舶动力设备是高能耗、高污染的机械，在其最佳性能下运行时，还能减少能耗、降低排放呢。”

第 8 章　数据挖掘在电信业中的应用

今天徐教授早早地来到了教室，看还有几位同学没来，就用手机看起 E-mail 来了。

过了一会儿，上课铃响了，他把手机放在讲台上，开始讲课："随着 3G 的广泛应用，4G 时代的到来，电信业发展面临着前所未有的机遇和挑战。运营商们都明白，客户占有量才是硬道理，于是他们千方百计地挖客户，但是客户也越来越'挑剔'了，不一定买他们的帐。于是，他们开始利用数据挖掘技术，对市场、对用户进行分析，进行科学化的决策。"

移动公司的梁总呼应道："电信业面临最紧迫的四大问题：第一个是市场分群，究竟客户是什么样子的；第二个精确营销，比如关联消费就是某一个用户用了你这方面的业务，此用户还会用其他什么方面的业务；第三个是新业务响应，你推出一个套餐、新业务，什么样的人来响应你；第四个是客户流失，什么样的客户会流失，为什么会流失，怎么预测他们的动向。"

徐教授喜笑颜开地说："梁总概括得很准确。接下来几节课，我们就讲解数据挖掘在这四个方面的应用。"

8.1 市场细分

“首先我们来讨论数据挖掘在电信市场细分中的应用，问一个问题，什么是市场细分？”徐教授提问道。

“市场细分，Market segmentation，是指营销者通过市场调研，依据消费者的需要和欲望、购买行为和购买习惯等方面的差异，把某一产品的市场整体划分为若干消费者群的市场分类过程。每一个消费者群就是一个细分市场，每一个细分市场都是具有类似需求倾向的消费者构成的群体。”电信公司的冯总回答道。

“市场细分应该有很多维度去区分，比如人口特征的划分、消费行为的划分等。最主要的是市场细分应该围绕着营销目标进行。”在冯总的基础上，华润万家的万总补充道。

徐教授又开始讲故事了：“说起目标的重要性，我想起了非洲猎狮的故事。话说某人去非洲打猎，找到当地著名的老猎人说要学打狮子。老猎手说你要打狮子，一定要讲究方法呀，第一要知道狮子在哪出没，比如，有草的地方，最好还有水；第二你要知道哪些地方狮子扎堆，数量多；第三你要知道什么样狮子不能打，比如怀了孩子、带着孩子的母狮子不能打，比如公狮母狮交配时节不能打，狮子在进食离它远一些，还有把狮子逼到绝境时不要拼命，诸如此类，最后老猎人强调：你一定要选准对象。”



大家都伸长脖子，等待徐教授故事的结尾。

徐教授继续讲到："此公认真记录下来，回去之后牢记老猎人的教导，做了充分的思想和物质准备，最后带着枪出去打狮子了。三天后，猎人欢天喜地回来了，大家纷纷涌向他的住处，去看他打回来的猎物。他把大家带到关狮子的笼子旁，众人定睛观看：笼子当中是一头既没怀孕，也不强壮、全身慵懒，性情温和的公狮子狗！——这是我在郊区公园树林中的一大群狮子当中抓回来的，容易得很，早知道的话根本不用带枪，此君大声向围观的人宣布。"

听了徐教授的故事，电力的刘总感慨地说："非洲猎狮的笑话告诉我们：不要由于关注于完美策略而忘记目标。我看过很多人，他们的计划从技术上讲完美无缺，你却很难从中发现他究竟想要获得什么？就像那个勇武的猎狮人一样，花费时间制定了计划，投入资金配备了装备，最终带回了一条狮子狗，错把狮子狗记作了他的狮子。"

徐教授点评道："很有意思的一个故事，作任何事目标始终是核心。电信企业的客户细分的目标可以概括为：通过对客户的人口统计特征、各业务消费特征等信息的有效挖掘和分析，制定适宜的营销策略、广告策略、促销策略、渠道策略等来实现公司更好的服务客户、增加企业的语音业务和各增值业务的使用量和收入的目的。"

汪部长思考了之后说出自己的观点："看市场细分，我认为核心是考察客户的行为模式。这类用户在电话使用上，服务使用上，是怎么样的行为模式。再比如客户喜欢到营业厅还是到其他渠道，所有这些东西综合起来，我相信对市场细分来说也是非常重要的"。

徐教授不断地点头示意，在教室中巡视，大家纷纷发表自己的意见。

"在我们电信业界中，最出名的市场细分的一个代表就是移动。其下三大品牌定位：全球通、动感地带、神州行，市场细分就做得很好。以动感地带为例，一般用户特征是年龄在 25 岁以下，在校学生，有一定彩铃和上网需求，容易接受新鲜事物。正是 2003 年动感地带的推出，进一步巩固了其行业老大的地位。"作为熟知电信业的资深人士，铁路的高局长侃侃而谈。

听完铁路的高局长讲述后，移动公司的梁总也说道："中国移动认定 25 岁以下的年轻新一代消费群体将成为未来移动通信市场最大的增值群体，因此，将以业务为导向的市场策略率先转向了以细分客户群体为导向的品牌战略，锁定 15～25 岁年龄段的学生、年轻白领，打造新的增值市场。事实证明，锁定这一消费群体来主打自己的新品牌，使中国移动动感地带品牌获得了巨大成功"。

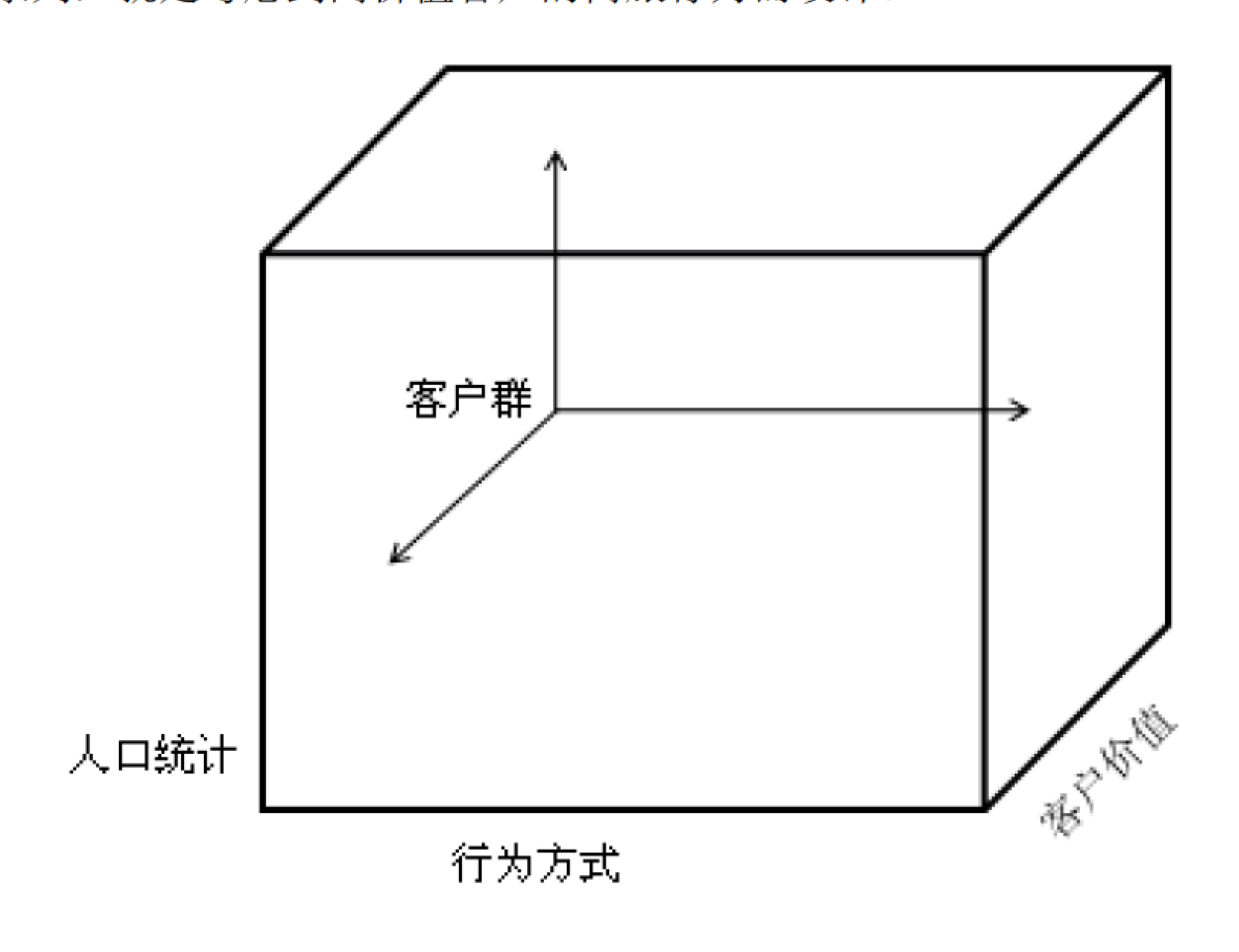
徐教授接过话题说道："大家都说得非常好，基本上很全面了，每个人都从自己的角度进行了描述。近几年，随着 3G 技术的发展，电信业发展日趋成熟。市场细分的工作也越来越具体，对不同客户的消费习惯、缴费方式、业务了解途径等均有其独特的特点，更多的时候需要针对某时间段、某类用户制定一个营销计划。"

台下一个学员说："徐教授，你还是外甥打灯笼——照旧（舅），结合一个例子给我们讲讲。"

徐教授回应道："在数据挖掘中，经常应用的方法是聚类。与分类模型有着本质的区别，聚类模型属于非预测模型（描述型模型）。聚类模型解决的问题是对用户进行分组（或者叫分群），特征相似的用户在一个组内，特征不同的用户分在不同的组。"

"那细分过程中一般考虑的维度是什么呢？"台下一个学员提问道。

铁路的高局长说："一般地，结合人口统计特征，是从价值和行为两维属性进行电信客户细分，从而实现了客户的人口特征—价值—行为的三维细分。并对客户价值—行为的一级细分的结果进行特征刻画，为营销策划提出参考建议。比如全球通的 88 套餐系列，就是考虑到高价值客户的商旅行为而设计。"

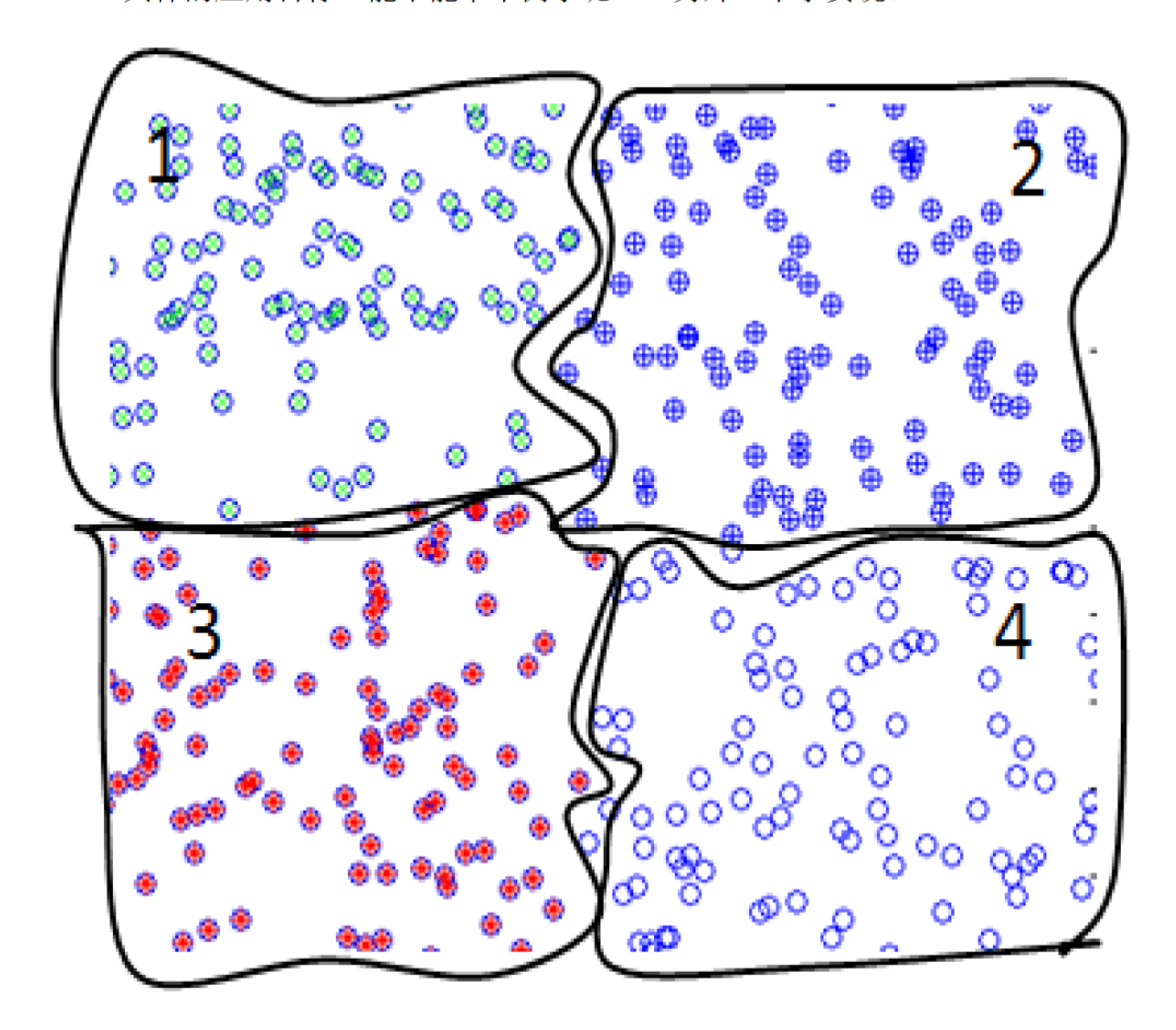


徐教授进一步解释道："从分群分类的角度来讲，有各种不同的纬度，我们注重的是什么呢？前面有人已经说过了，主要按照用户的消费行为，即他/她打电话的具体行为。如果用统计的方法（年龄、性别、地区等），这个很快就可以做出来。但是，这个太简单，我们要用更多的变量，来做行为方面的分群分类"。

台下一个学员问道："徐教授，在聚类的过程中，需要注意什么呢？"

徐教授解释说："聚类模型不需要目标变量，只需要给定自变量，聚类模型就可以自动地对用户进行分组，输出每个样本对应的组编号。选择聚类所需的变量是构建聚类模型最关键的工作，变量的选择往往取决于应用的目标要求。"

"具体的应用目标？能不能举个例子呢？"另外一个学员说。



徐教授说："以电话语音业务来说，想了解目前客户的语音分布情况。就可以用通话的相关数据变量（比如本地主叫、本地直拨长途、漫游主叫、漫游被叫、通话时长、通话次数等），利用聚类技术把客户刻划成4个类型：呼入为主、长途强势、IP突出、夜间积极。"

一个学员问道：“徐教授，那有了这个市场细分结果，怎么制定营销策略呢？”

徐教授：“这个方法就很多了，比如长途强势组，可以针对此组客户推荐省内‘两城一家’和省际定向长途包；再比如针对夜间积极的组可主动推荐忙闲时价格差异化的语音套餐；针对呼入为主的客户组（这部分以老人为主）可以推荐免月租的套餐或者进行家庭宽带业务捆绑等。”

铁路的高局长感慨道：“听了很多次关于市场细分的讲座，像徐教授这么专业、又这么懂业务的人真是很少，今天学习了很多。相信进行了良好的客户群体细分之后，必定能帮助我们电信业更好地进行客户关系管理，提供更优质、更专业、更贴心的服务”。

8.2　精确营销

马上就要上课了，大家都被大屏幕上的题目“数据挖掘在精确营销中的应用”震住了，私下三五成群的窃窃私语。

“精确营销是个什么概念啊，头一回见。”有一个学员说道。

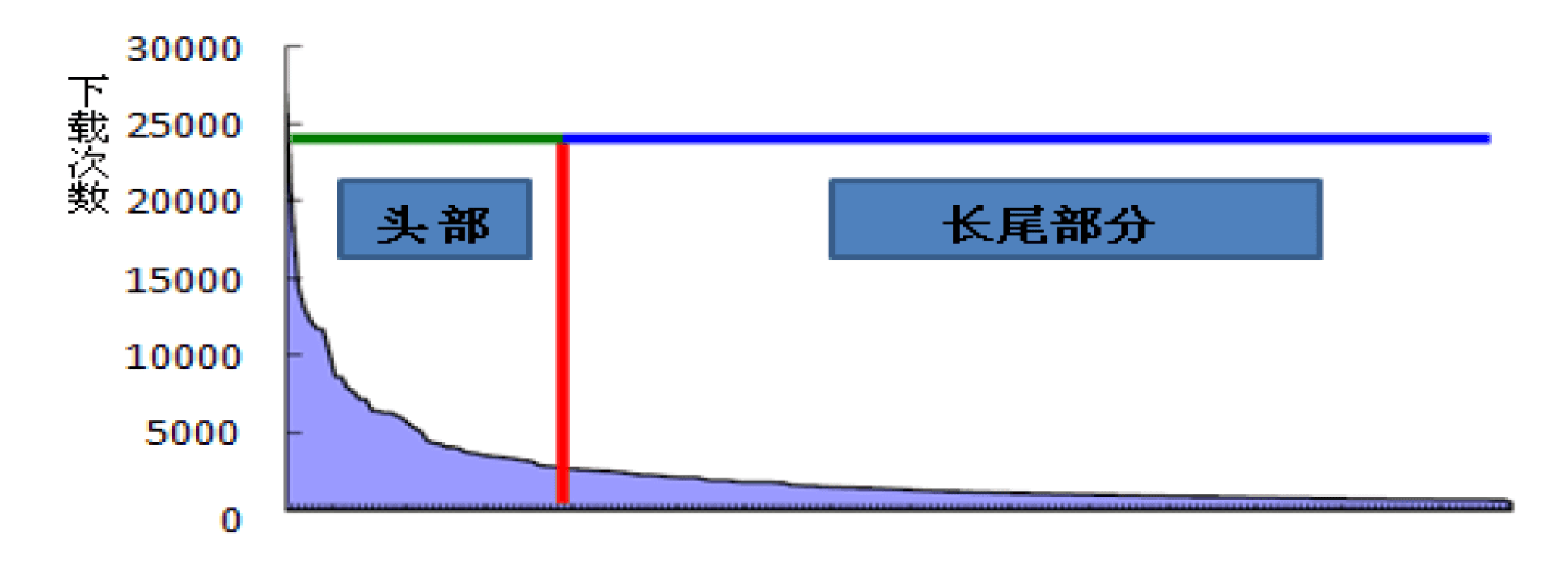
“大众营销我倒是听说过，但是精确营销也不知道是什么呢。”另外一个学员也不知道。

徐教授这时走进了教室，说道：“营销学中有著名的‘二八定律’和‘长尾理论’。‘二八定律’说的是企业应该关注重要的人和重要的事，即重点针对创造 80%利润的 20%的客户做营销。”

“是这么个理，应该抓主要问题嘛。那‘长尾理论’说的是什么呢？”台下的华润万家的万总急切地问道。

徐教授说道：“以移动电话运营商的彩铃业务为例，可以供客户下载的歌曲有上万首，这样用户便面临着无限的选择，而其中的每一首歌曲都有可能被用户下载，尽

管绝大部分歌曲下载的需求和实际下载量并不高，但这些处于长尾部分的下载量占总下载量的比例加在一起却可能超过正态曲线分布中处于头部位置主流歌曲的比例，也就是说那些不流行的、占绝大多数的彩铃相对于流行的、少数主流彩铃所创造的收入和利润要更多，这就是目前颇为流行的长尾理论。”



听完徐教授的举例介绍，大家都点头示意明白。

接着，徐教授说：“长尾理论告诉我们，不仅要关注处于传统需求曲线上那个代表畅销品的头部，更要关注所谓冷销品的长尾部，这就需要我们要更深入地研究目标客户群体和个体之间的需求差异。精确营销正好能帮助我们更好地分析和研究目标客户群体和个体需求”。

徐教授引入到正题了，大家也更加专心了。

徐教授说：“市场竞争日趋激烈，客户出现了日趋个性化的偏好与需求。面对客户的多样化、层次化和个性化的偏好与需求，传统大众化的营销就失去了优势。大家上课前的讨论我听见了，其实大众营销和精确营销一个重要区别就是：精确营销的推广销售群体是有针对性的目标用户，而传统营销则面对的是所有大众。”

“现代营销之父菲利普•科特勒先生曾指出：促销费用的大部分都打了水漂，仅有 1/10 的促销活动能得到高于 5%的响应率，而这个可怜的数字还在逐年递减。徐教授，这样针对性的选定目标用户进行营销，就可以节省广告宣传费用，收益能提高不少吧？”台下一个学员小声地问。

徐教授："假设某企业有客户群 25 万人，希望对他们做一次邮寄的促销活动，每一个用户邮寄成本为 1.5 元。如果客户对促销活动响应了，平均能带来 200 元的利润。对 25 万用户全部邮寄，如果响应率在 1%左右，那么收益了 125000 元。"

大家都惊讶广告支出费用的庞大，想着若是换了小公司，岂不是企业不能承受之重。

徐教授接着讲到:"通过精确的目标用户筛选,选择 2.5 万用户(取总用户的 10%),如果响应率达到 5%（取原来的 5 倍），那么收益为 212500，比对全体用户邮寄的收益提高了 87500 元。"

鼓风动力集团的王总提问道："徐教授，通过这个例子我们都看到精确营销在节约营销成本、提高利润水平上无疑比传统营销更具优势。但是这个例子成立前提有一个假设，就是筛选 10%的用户、响应率要达到 5%（是原来的 5 倍），精确营销怎么做到这一点呢？"

鼓风动力集团的王总一问之后，大家也都意识到了这个问题，都在等待徐教授的解答。

徐教授不慌不忙地说道："这问题提得相当好，也正是我要给大家讲的。精确营销是一个基于数据分析的量化过程，对用户使用行为和偏好的精准衡量和分析，从而实现在合适的时间、合适的地点精确推荐给合适的人。而传统营销更多采用市场调研方式了解客户消费行为及偏好，定性分析和主观因素要更多，而且客户某些潜在的需求和间接的偏好是无法通过调研得出所有答案的。"

刘经理激动地说道："哦，大众营销好比古代打仗的时候，知道有敌人，但是了解敌人不够透彻，乱射箭，命中率就比较低。精确营销呢，就好比熟知敌人的特性，锁定目标进行攻击。虽然发出的箭不多了，但命中率反而大大提高了"。



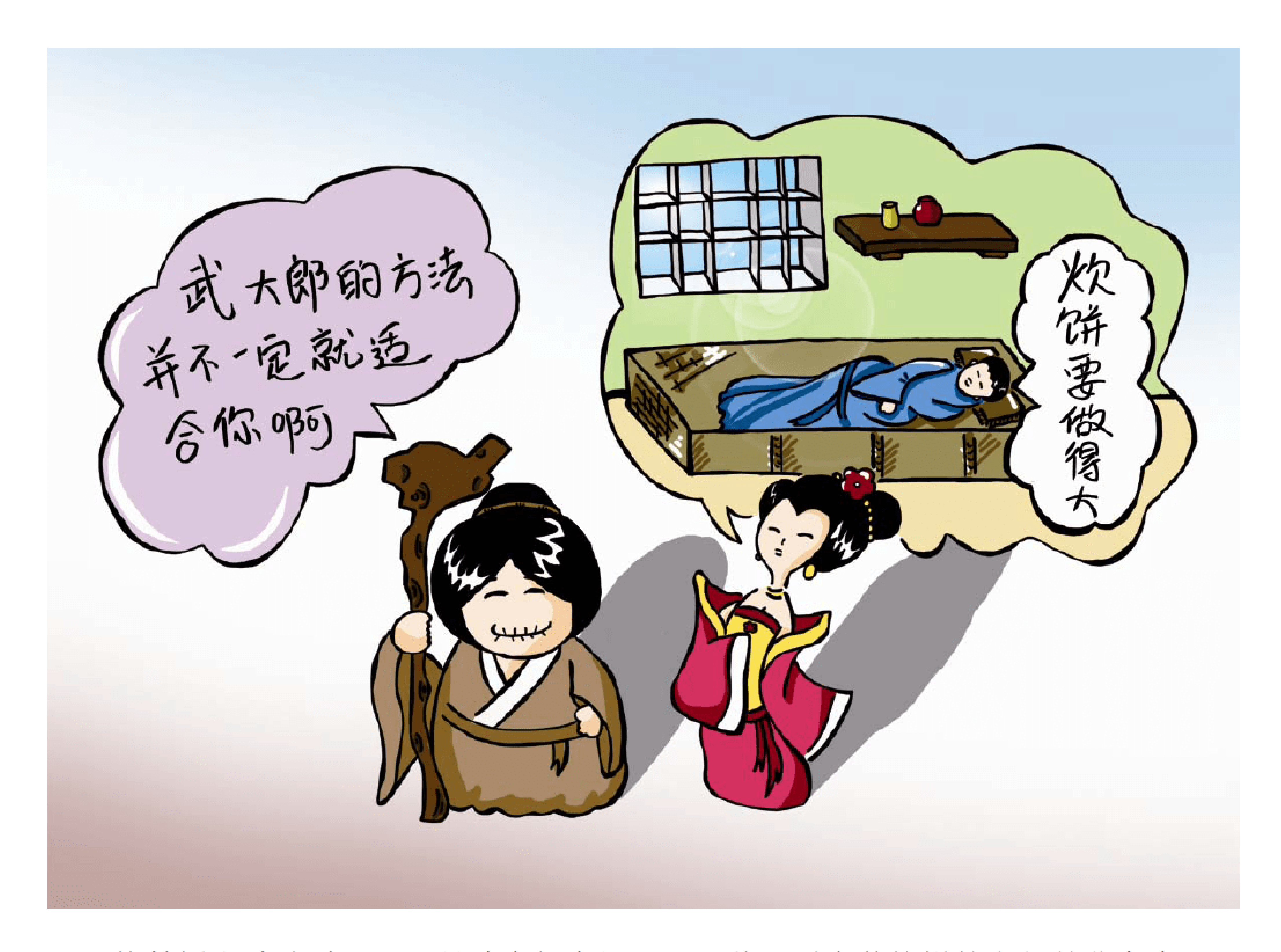
台下一个学员打趣道："刘经理真是有意思，从某种程度上来说，营销还真是像打仗"。

徐教授："我们都知道武大郎死得很惨，但真正知道武大郎临死前最后一句话内容是什么的人并不多。事实上，经过多方考证，最终发现武大郎对潘金莲讲的最后一句话是'炊饼要做得大！'至于为什么要做得大，潘金莲并不明白。"

大家都纳闷徐教授怎么突然讲起水浒了，还是个外传，都兴致勃勃地等待下文。

徐教授继续说道："潘金莲听了大郎的临终嘱咐之后，以后的炊饼都做得很大，不过她发现不管小炊饼还是大炊饼都不影响她的销售业绩，因此很疑惑，就去问王婆为什么大郎死前说要把炊饼做得大。王婆不愧老江湖了，她告诉金莲：大郎的个子矮，把炊饼卖给客户的时候，怕那些人弯腰够炊饼，如果做得大一些，客人不用弯腰也可以够到大郎的炊饼。"

"原来是这么回事。"台下的学员恍然大悟。



徐教授幽默地说：“可是武大郎忘记了，那些买潘金莲炊饼的人却并非高个子，根本用不着大炊饼这一招，而且找潘金莲买炊饼的人更多关注的是她的脸，而不是手中的炊饼。”

大家都被逗乐了，可是大家都还不明白徐教授讲的这个“大郎遗嘱”有什么用意。

徐教授揭开谜底说：“‘大郎遗嘱’的笑话告诉我们：对你来讲，精准定位并能恰当把握的群体，却未必是别人眼中具有定位的群体。所以，做精确营销的时候，如何确定一个大家都认为是有价值的目标对象呢？”

原来徐教授讲“大郎遗嘱”故事的“醉翁之意”在这呢，大家默契地回答道：“数据挖掘！”

徐教授特别用一句东北方言说：“哎呀妈呀，大家都老有默契了。”

随后，他顺着刚才的话题继续说道："精确营销解决的问题是：哪些用户是某个产品或者营销活动的目标用户？每个用户最适合被推荐的产品是什么？数据挖掘正是通过对客户消费行为数据和历史规律的挖掘与分析，进而可以找到目标用户的特征，实现以客户为中心的精确营销"。

"徐教授，那精确营销的时候，经常用到的数据挖掘手段有哪几种呢？"台下一个学员问道。

徐教授解释道："在精确营销领域，数据挖掘范围很广，比如分类、聚类、关联等。今天我们就学习一下关联在精确营销中的应用。关联模型主要可以解决两大类问题：一是对用户进行商品推荐，即交叉销售问题；二是哪些商品在一起销售更好？即捆绑销售问题。"

"交叉销售？捆绑销售？今天上课新名词还真不少。"台下一个学员说。

徐教授回答："对，交叉销售，就是发现客户有多种需求，通过销售多种相关产品或服务的营销方式。比如，某碳酸饮料厂商把自己的饮料和薯片捆绑在一起销售，年轻人在吃薯片的时候，喜欢喝碳酸饮料，薯片降价，自然会促进这种饮料的销售。"

移动公司的梁总说："徐教授，我明白了。前面你讲过：关联模型主要解决的问题是研究产品购买的关联性，即买 A 产品的同时是否会对 B 产品也感兴趣。你这个案例中交叉和捆绑销售的应用就是发现购买饮料的同时购买薯片的可能性比较大。"

华润万家的万总说："对，我也记得前面说过：关联模型又叫购物篮分析，在超市购物时一个购物车中往往会放多种不同的商品，通过对大量的购物车进行分析，这些商品之间可能会存在众多意料之中或意料之外的关联性。"

徐教授进一步阐述："你们都说得不错，关联模型中度量两个产品关联性强弱主要用三个指标：支持度、可信度和提升度。考考大家对三个指标的认识，谁先来说说什么是支持度？"

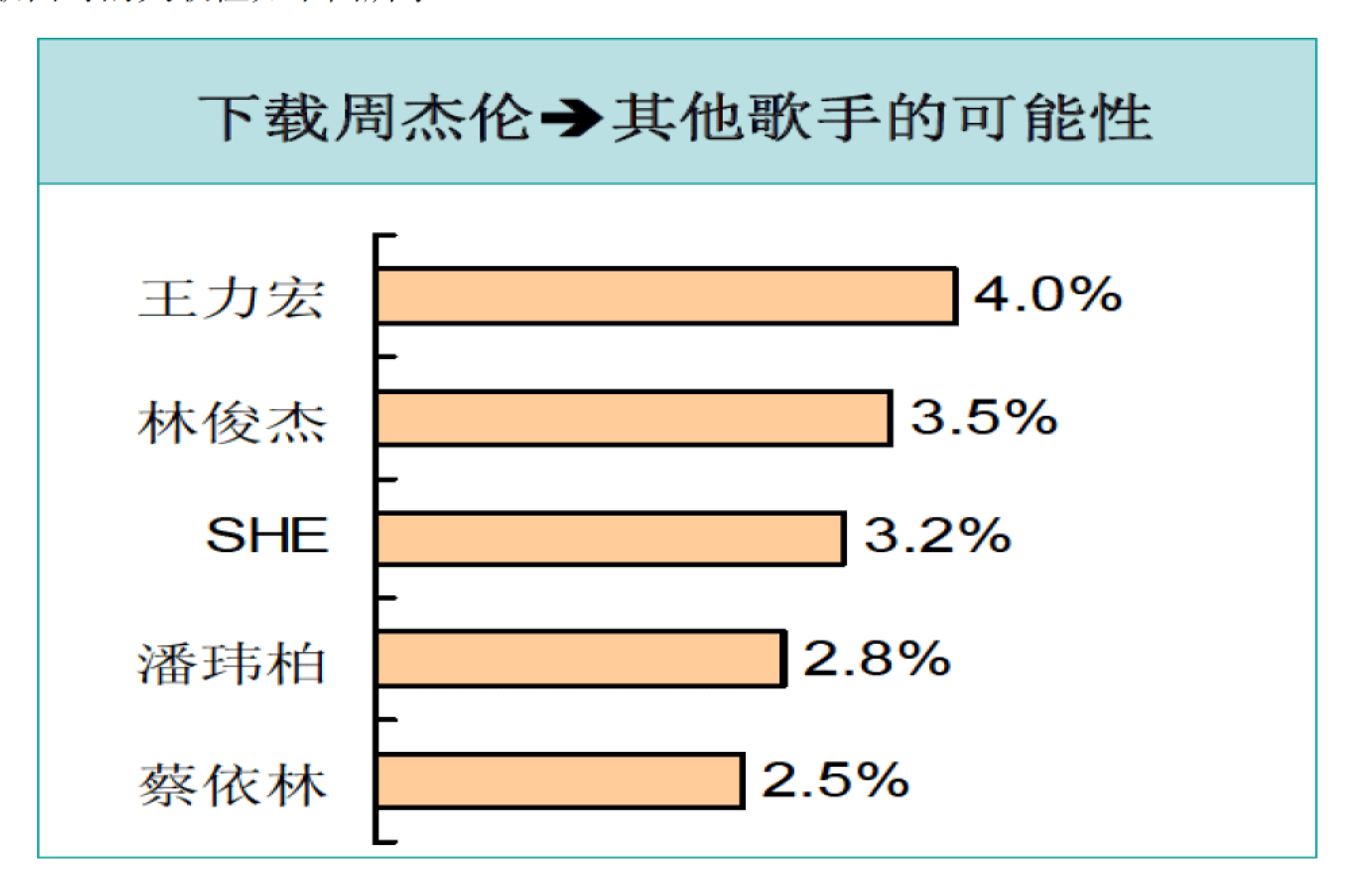
“支持度，Support，就是表示 A、B 同时购买的人数占总购买人数的比例。支持度越高，表示商品同时购买 A、B 的人数越多，这两个商品越主流。”台下一个学员回答道。

徐教授说：“回答得很正确，可信度呢？谁来讲讲？”。

“可信度，Confidence，表示在购买 A 商品的人中同时购买了 B 商品的比例。可信度越高，表示购买了 A 商品后再购买 B 商品的可能性就越大。”台下的另外一个学员回答道。

华润万家的万总举手说道：“提升度，lift，可信度除以总用户中购买过 B 商品的用户占比。提升度越高，表示购买了 A 商品对购买 B 商品的影响度就越大，也即他们之间的相关性就越强。”

徐教授点评道：“呵呵，都有抢答的了，回答也很正确，加十分。以电信运营商的彩铃为例，我们把歌曲或者歌手当做商品来研究，用户在订购歌曲或者某个歌手的歌曲时的关联性如下图所示：”

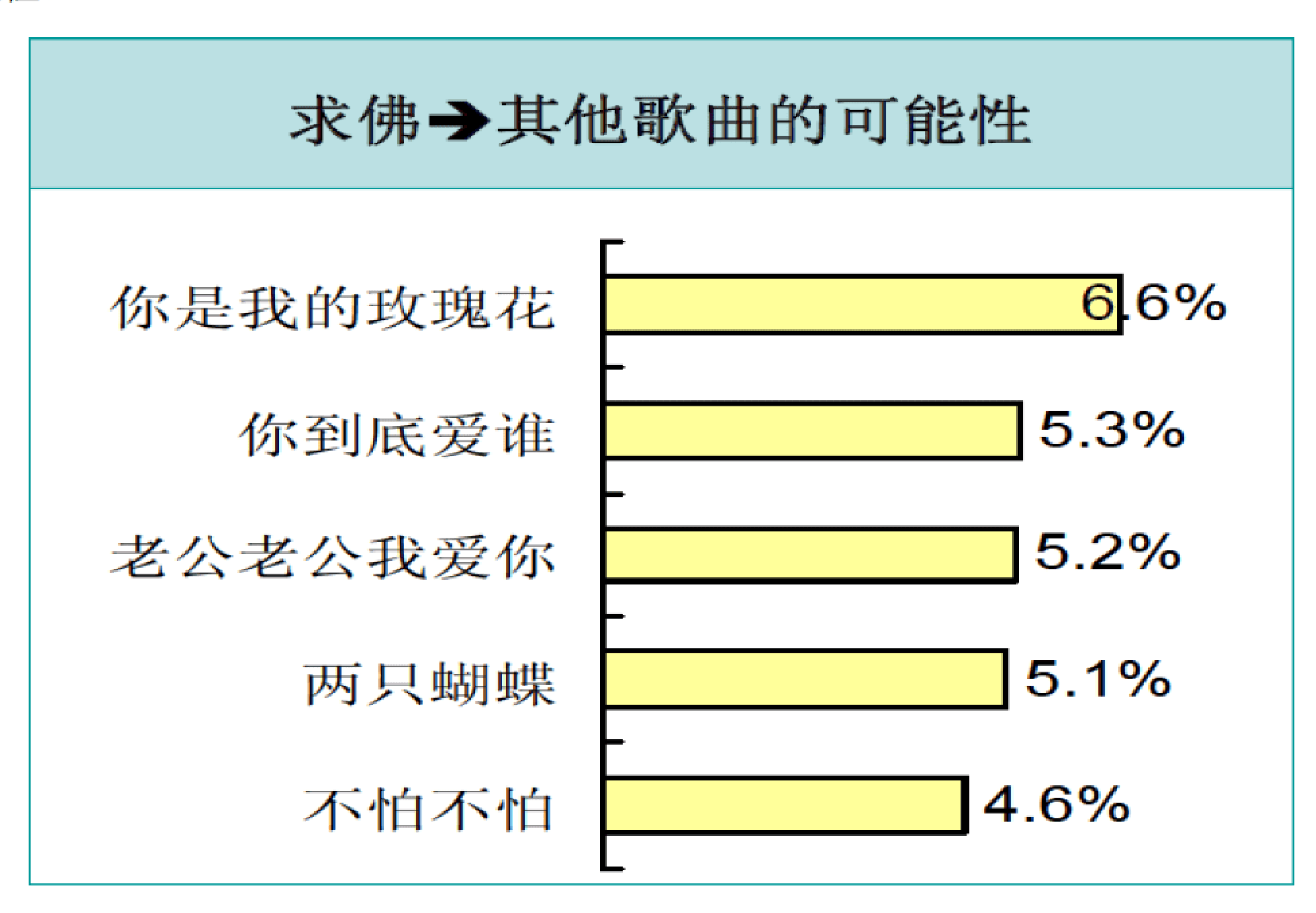


徐教授针对关联的歌手结果解释道：“从上图可以看出：下载过周杰伦歌曲的用户中，还下载过王力宏的比例最高，林俊杰次之。因此可以针对下载过周杰伦歌曲的用户推荐王力宏或者林俊杰的歌曲交叉销售。”

刘经理说：“这个结果我觉得有一定的准确性，就拿我女儿来说，整天在家念念叨叨的就是周杰伦、王力宏、林俊杰，有演唱会这丫头必然去凑热闹。”

台下另外一个学员说：“哦，根据这个关联结果，电信营销彩铃的人员就可以把周杰伦、王力宏、林俊杰的歌曲捆绑在一起打折销售给客户啦。”

徐教授：“上面的关联彩铃是基于歌手的，下面我们一起看看关于歌曲之间的关联性。”



电信业的铁路的高局长激动地说：“徐教授，看见这个图我终于明白了。根据基于关联性的结果就可以进行捆绑销售：将‘求佛’、‘你是我的玫瑰花’、‘你到底爱谁’等歌曲捆绑销售 3 元。”

徐教授点头后讲道："对，关联结果应用在捆绑销售中的时候，还有一个重要的原则：目标顾客的一致性。也就是说，捆绑在一起的几种商品，其主流的消费者群体应该是一致的。比如年轻人多喜欢周杰伦等流行音乐，中老年人喜欢比较怀旧、古典、民族音乐，那么在推送捆绑彩铃时若是胡乱推荐恐怕就不会有良好的促销效果。"

听到这里，大家都很受启发，原来数据挖掘在精确营销里面可以真正做到以客户为中心，不吹一点牛。

电力刘总的一席话道出了大家的心声："数据挖掘技术作为支撑精确营销的重要手段，随着企业对精确营销认识的提升和需求的加强以及其本身算法的不断完善，必将在未来的营销领域中发挥强大的作用。"

徐教授也鼓励在座的学员："随着商业竞争的日益激烈和信息技术的突破性进展，营销界正在爆发一场意义深远的革命，无论是营销理论还是实践都面临着一种结构化的转型：从传统的、大众的和粗糙的方法跃变到深度化、细分化和精确化的模式。任何公司要在这场革命中占领先机，都必须坚定不移地聚焦客户，并开始向精确营销转型。"

8.3　业务响应

徐教授："对于电信企业来说，竞争已不仅仅来自行业内部，终端企业、互联网企业等都在动摇其价值链的核心地位，使运营商的管道化趋势日益明显。"

铁路的高局长认同地讲到："是的，现在竞争越来越激烈。面对严峻的形势，电信企业需要重新定位。未来电信企业除提供最基本的语音、短信、彩信等通信服务外，更重要的是提供差异化的专有服务和开放电信能力搭建差异化的数字内容集成平台。"

徐教授："与语音业务发展进入饱和期不同，数据业务近期取得了快速发展，运营商如果想保持其原有的收益，就不会甘心沦为管道商，必须向数据业务转型。"

台下的电力的刘总说道：“推出业务，关键是看需求，需求是源动力。”

移动公司的梁总也表述道：“是的，市场调研能帮助我们了解一些业务的需求。此外，在业务运行之前，没有条件预演的情况下需要预测市场反应。比如你推出一个套餐、新业务，什么人来响应你。”

华润万家的万总笑着说道：“想起关于市场反应的一个笑话：巡回调查几个星期后，推销员向上司报告：市场上只有两种反应。上司问哪两种，推销员说：‘滚出去’和‘住口’两种。虽然是个笑话，但是我们从一个侧面可以看出：想知道推广业务的市场反应还是有一定难度的。”



徐教授说道：“这就是我们本节课的核心内容：业务响应方面，数据挖掘能做些什么？”

看了看周围，汪部长说：“徐教授，我看大家的意思都等着你讲一个实际例子呢。”

徐教授摆了摆手，接着说："好，那就顺大家的意思。假设现在某电信运营商正准备推销某种增值业务，需要寻找有购买潜力的目标用户特征，即哪些客户可能会对这个增值业务响应积极。"

台下一个学员说："徐教授，您就别卖关子了。目标我们已经明确了，就是确定营销某增值业务的响应用户群。在这之前，需要准备些啥？"

台下的另一个学员说道："我知道一些，一般电信中用户的行为数据包含：电话使用的方式、服务使用的种类；用户的人口统计数据包含：年龄、性别、地址；细分中可能还需要的一些其他数据，比如帐账户设立时间、网络质量、客户关怀、级别等。"

徐教授赞许地说："一看就是内行，简洁到位。数据有了，我们就该想着用什么方法了。在建立客户业务响应模型时，应用到数据挖掘手段主要有分类技术，这里我们就用分类来说说。"

"分类？这之前我们接触过。当时学的分类的应用实例，但是时间久了记不清楚了。"台下一个学员抓了抓自己的头说道。

鼓风动力集团的王总得意地说道："这个我记得，分类原理我比较清楚。分类是数据挖掘应用最广泛的应用之一，属于预测性模型。分类模型解决的问题是对类别未知的用户进行预测，以判断其属于哪个类别的概率比较高。"

徐教授肯定地点了点头，环视了一圈下面学员，期待更多人发表自己的意见。

移动公司的梁总也不甘示弱地说："分类模型的构建需要一个类别已知的历史样本——训练样本。由于训练样本中每一个个体的类别都是明确的，因此可以通过分类的算法找出能显著区别不同类别的典型特征，这些特征就是分类模型的结果。通过训练样本找出来的特征，对新样本进行预测，以判断满足不同特征的用户属于不同的类别。"

徐教授说："大家都讲得非常好。决策树是分类模型中最常用的方法之一，具有预测精度高、预测结果稳定性好、结果易理解等优点。除了决策树之外，Logistic 回

归、神经网络、判别分析等方法也可以构建分类模型。这里我就给大家展示一下决策树的分类方法。”

台下一个学员问道：“徐教授，按照数据挖掘的流程。现在的步骤应该是数据预处理了，那么在数据准备选取中，需要注意什么呢？”

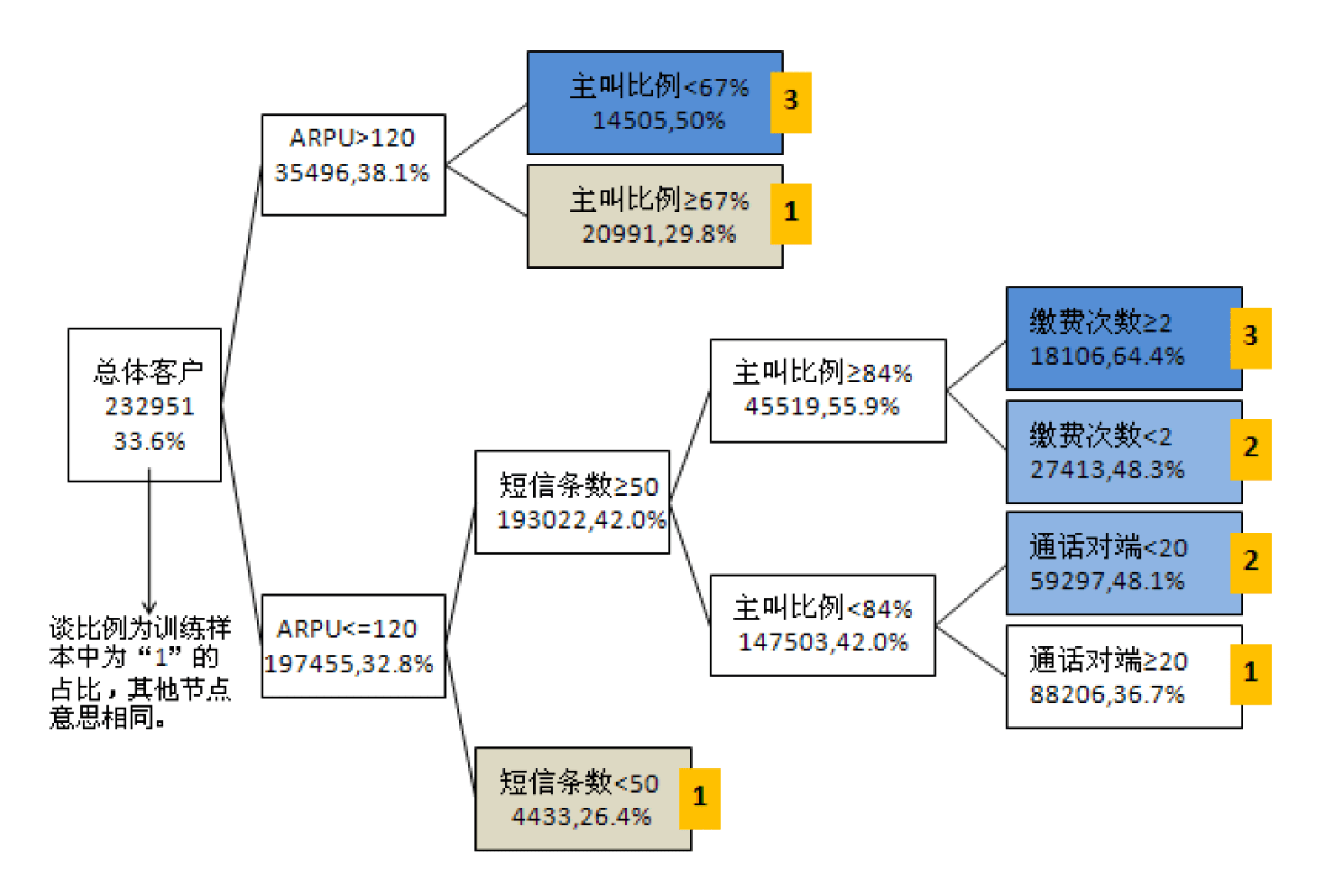
“强调一点：保证数据的平衡型，合理分配训练样本和测试样本。比如这里可以选取 23 万客户，在训练样本中有 33.6%的用户已经订购了号簿管家这个增值业务（定义为目标变量取值 1），其余 66.4%的用户均未订购（定义为目标变量取值 0）。”徐教授回答道。

“号簿管家这个业务是什么？”台下一个学员提问道。

移动公司的梁总解释道：“号簿管家是中国移动推出的一个专业服务于移动电话用户的通讯录业务，通过 Web、WAP、短信、SyncML 等多种方式，为移动电话用户提供最为便捷、安全、有效的个人地址服务。可以通过手机短信、WAP、PC 等多种方式对个人地址本进行维护、管理、查询，是如影随形的‘活的通讯录’。它还提供了短信群发、电子名片册、短信收藏夹、日程管理等增值功能。”

徐教授接着说：“采用决策树方法构建分类模型，可以看到满足‘ARPU>120’并且‘主叫比例<67%’特征的用户中有 50%订购了号簿管家，显著高于总体中的 33.6%，因此可以认为满足该特征的用户购买号簿管家这个增值业务的可能性比较高。”

看大家全神贯注，徐教授接着说：“同样，我们还可以看到‘ARPU<=120’并且‘短信条数>=50’并且‘主叫比例>=84%’并且‘缴费次数>=2’的用户购买号簿管家业务的可能性会更高，达到了 64.4%。”



最后，徐教授说道：“现在我们已经从上图看出来针对新推出的号簿管家业务，第三组的购买意愿最强大，第二组的相对较强，一组的意愿最小。大家都是聪明人，接下来该怎么做相信都有各自的十八般武艺了。”

刘经理第一个献策道：“可以获得每个客户分组中所有客户或部分客户的名单进行呼叫。”

华润万家的万总接着补充道：“可以灵活地对形成的各客户分组进行宏观观察和微观细分，就是追踪和监视分类结果。”

“可以借助计算机程序动态观测客户行为的变化及其所属客户细分群体的变化，测算前后的收入变化。”鼓风动力集团的王总落实到企业最关注的利益收入上。

结合现在，展望未来，移动公司的梁总说：“已经推广的数据业务，准备推广的数据业务，在未来的 3G 平台上，这些业务会更加丰富。通过上述对客户行为模式的分析，能够更好地划分客户并进行针对性产品设计和市场营销。”

8.4 客户流失分析

徐教授今天直入主题："开始上课了。我们先来讨论一个问题：新客户获取和老客户保留。欢迎各位学员就这两个问题分享自己的看法"。

航天的黄主任率先表态："我觉得新获取客户比较困难：招揽一个新的客户，我们销售员需要笑脸相迎，百般讨好，有时候客户却视而不见；在说服一个新客户的时候，销售员需要费尽口舌，做足工作，但有时客户需要另行比较。"

李经理则持相反态度："打江山容易，守江山难。发现老客户要流失并采取措施留住一个老客户，这不是一件容易的事情。"

电力的刘总说："据第三方调查研究表明：对企业来说，新获取客户成本是挽留客户的成本的 5 倍！"

鼓风动力集团的王总说："我觉得做好老顾客的保留，有助于新客户的获取。大家都知道口碑效应，通过亲朋好友的推荐成功率比企业自己去推荐，成功概率要高很多。"

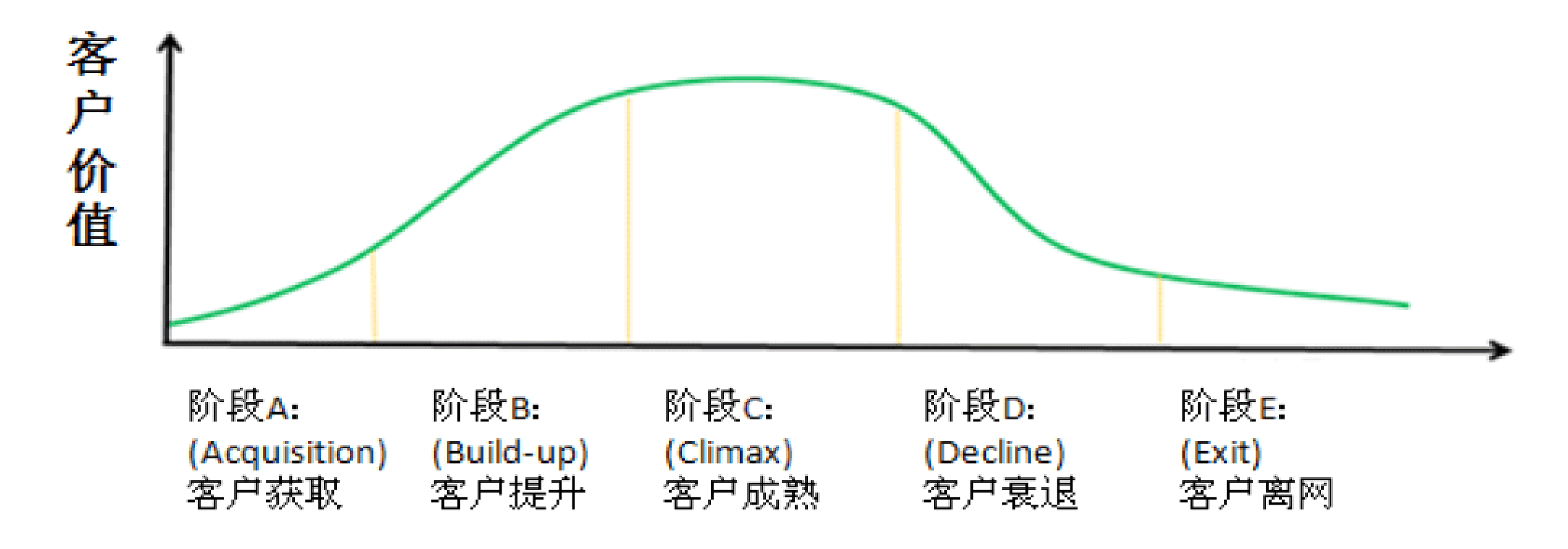
鼓风动力集团的王总说完关于新老客户获取看法后，赢得到了大伙的一致认可。

徐教授接着说道："客户保留对企业生存发展的重要性是毋庸置疑的。以美国无线业流失的数据来看，其国内客户流失率，一个月是 2%，全年就是 25%，非常巨大。那么在面临如此高的客户流失率的情况下，作为运营商应该怎么样处理这样的问题？整个客户流失管理，已经成为运营商非常关切的问题。"

铁路的高局长感同身受地说："客户流失是电信行业永远会存在的一个问题，无法避免。当一个新的电信产品出现的时候，在早期会吸引到部分客户（比如一些'发烧友'或'尝鲜族'），要是营销效果差的话，在短暂的用户增长过后，就会有一个用户数量的下降、流失。"

徐教授说："在客户生命周期图中，我们可以明显看出：在产品上市初期的一段时间（阶段 A 和阶段 B）用户呈增长趋势；在用户快速增长之后是相对的成熟期（阶

段 C），数量稳定，客户价值比较高；接着不可避免地迈入阶段 D 和阶段 E，伴随客户总数整体下降，客户为企业提供的价值也跟随着下降。”



移动公司的梁总问道：“徐教授，既然客户流失对电信行业来说是必然存在的一个问题，且不可避免。那么面对客户流失，我们能做些什么呢？坐以待毙肯定不行，这与企业的追求目标是相悖的。”

徐教授直奔主题解释道：“什么样的客户会流失，我们应该怎么预测他们；他们为什么会流失，我们应该怎么挽留他们，哪个部分的客户是我们应该留下来的？诸如此类的问题都是数据挖掘可以在客户流失方面做的工作。”

这时，铁路的高局长插话：“一般电信业务客户流失考虑的都是可控因素引起的客户流失分析，由于不可控原因（比如用户死亡）造成的流失是不考察的。比如最近某客户没工作了，自然就不想用电话了，座机可能都拔掉了。再比如某客户的生活地点发生了变化（以前是生活在沈阳，后来去了深圳），那肯定手机要换掉了，这个用户是挽留不了的。”

刘经理提问：“徐教授，刚才高局长说到考察可控因素引起的客户流失，那针对这类情况的客户流失，一般考察哪些维度呢？”

徐教授温和地解释道：“流失也可以有各种不同的角度，我给大家简单举几个例子。比如找出其人口特征，如业务使用情况及入网时间特征，研究表明入网时间两年以上的用户比较稳定，达到五年以上的流失率更低。”

喝水停了停后，徐教授接着说：“我们可以找出他/她的消费特征等，比如网间跳转造成客户流失，考察其很重要一个因素就是呼叫转移，重点研究其交往圈。还有一些特殊考察，比如疑似双卡用户的甄别分析，这类客户一个人用 2 个号码，比如用一个移动号码，用一个的联通号码，他们的流失率被证实是非常高的。”

鼓风动力集团的王总祈求道：“徐教授，我对这方面内容非常感兴趣，能不能详细地通过一个例子来给我们说明一下呢？”

徐教授答道：“电信客户流失分析最基本的方法是分类和预测方法，这里举个决策树方法的例子。它的优点在于它可以生成可以理解的规则，计算量相对较小，可以处理连续字段，并且可以清晰地显示哪些字段比较重要。”

铁路的高局长也谈起自己的经验：“很多客户流失的预警，要注意怎么把流失客户能够剥离出来。这里并不是针对一个人做活动，而是要看到，哪个群体的人的流失概率非常高。”

徐教授接着讲：“通过业务经验，针对客户产品拥有情况、入网时长、服务开通情况、优惠套餐信息、客户投诉情况、语音通话、月租费、优惠费用、缴欠费信息等进行筛选后，入选进入逻辑回归模型参数，将是否流失作为目标函数。”

影响客户流失因素	时间窗口
当月通话时长降幅	以 11 月份的拆机用户为训练目标，以 8～10 月份的数据为训练资料进行建模，之后用 9～11 月份的数据对 12 月份的拆机用户进行检验测试。
当月通话次数降幅	
当月消费额降幅	
当前欠费情况	
申请停机状态	
月均优惠费用	
当前优惠捆绑到期情况	
在网时长	
月均投诉次数	
使用增值业务种类数	
拨打竞争对手网用户的通话时长占比	
是否使用家庭宽带	

华润万家的万总求证说："徐教授，根据之前关于决策树的学习，加上刚才模型所使用的变量，最后模型能计算出的有：每一个用户是否会流失，流失的概率有多少。不知道是不是我理解的这个意思？"

徐教授说："很对，最后的流失预测值我们是反馈到一个客户数据库里面，每天更新一次，然后生成一个专门的客户流失清单，业务人员一打开就知道这个客户会不会流失，如果是 0 就不会流失，如果是 1 就可能会流失。"

铁路的高局长接着徐教授的话题说："现有国内的流失分析，多以'月'为单位，隐藏了潜在的流失消费特征。徐教授以'日'为单位进行分析，提高了客户流失预测的准确率。将这些信息写回到数据库的个人信息里面去，业务人员就可以根据流失预警级别进行关注了。"

汪部长问道："真是智能化，这里一个重要环节就是模型的预测效果怎么样？我们也看出来了，若是模型效果不行恐怕会弄巧成拙。徐教授，模型准确性这方面怎么评估呢?"

徐教授说道："一般地，决策树评价指标有三个，提升率、查全率和命中率。这三个指标越高，表示模型效果越好。此外，加上一个时效性的评估指标更合理些（在电信流失预测时，时间窗口挪动对预测准确性有一定影响）。这个想法是来源于当初大家讨论日本地震预测比国内高明，体现两点，第一是准确率，第二是时间提前量，比如我能预测到肯定地震，但只能在地震前 1 秒内，基本没意义，评价预测需要加上时间就更为合理。"

评估指标	决策树模型评估指标解释
提升率	客户的命中率/不使用模型时的流失率
查全率	被准确预测为流失的客户/样本中流失总数
命中率	每组中实际流失的客户/全组客户总数

铁路的高局长认同地补充道："徐教授说得很有价值。实践证明，数据挖掘模型需要不断调整，模型维护工作的简化是未来研究的一个方向。现有的指标体系还需要不断完善和深入。对于电信客户的流失分析，数据业务的蓬勃发展、3G 的到来等均会增减相关指标。"

接着，徐教授询问了一下学员关于防止客户流失的策略制定。在前面内容的铺垫下，大家纷纷发表意见。

电力的刘总说："关于客户流失管理策略制定，比如可以通过赞助一些演唱会，票不卖，只有运营商的用户才能拿到这个票，这个是赞助性的活动，客户喜欢留在你的品牌下。"

李经理也支招道："再比如俱乐部，高尔夫俱乐部，优先客户资格方案，金卡、银卡、钻石卡，用来挽留一些高端用户。"

汪部长补充道："确实是，可以理解，若是你搞一个积分奖励计划，对于高端客户并不一定有大家想象的那么好。你想一年打几万块钱的电话换回来一点点东西，真正的高端客户是不会很在乎这个积分奖励的，转向普通客户中经常参加积分兑换活动的人效果肯定绝妙。"

下课铃响了，但是大家还意犹未尽，仿佛每个人都是电信业的工作人员了，在那里商讨并献计献策。

第 9 章　Web 数据挖掘

今天，徐教授一进教室看大家都到了，就开始了徐氏开场，“同学们，大家都是老网民了，但如何衡量一个网站运营是否成功？网站中哪些内容是人气最旺的？主要访客是哪些人？什么原因吸引他们前来？要回答这些问题是不容易的，因为影响因素太多了，但这些问题都属于今天我们讨论的互联网和电子商务的数据挖掘的范畴。”

9.1 Web 数据挖掘概述

彭部长呼应道："现在电商间价格战打的很火，也听了这么长时间课，早就感觉他们一定用了数据挖掘技术，但就是不知道他们怎么用的。"

徐教授微微一笑说：你是个有心人啊！我们已经知道数据挖掘是从数据中提取新的、潜在有用的知识的过程，将数据挖掘技术与 Web 技术结合起来，从互联网信息中发掘出有用的模式在当今这个互联时代就显得非常重要。总体上说 Web 挖掘可以分为 3 类。

- 第一类是 Web 内容挖掘，是从文档内容或其描述中抽取内容，主要包含文本挖掘（包括 text、HTML、XML 等格式）和多媒体挖掘（包括 image、audio、video 等媒体类型）；
- 第二类是 Web 链接结构挖掘，是从 WWW 的组织和链接结构中推导知识；
- 第三类是 Web 使用模式挖掘，是从 Web 的访问记录中抽取感兴趣的模式。



李部长一听彭部长被表扬了，就也迫不急待的说："现在网购时，电商给我推荐的那些产品，是不是就是属于使用模式挖掘啊？他们通过我的浏览记录，知道我可能会对哪类商品感兴趣，就会给我推荐"

徐教授答着说："嗯，很对，李部长是个会思考的人，实际上 Web 挖掘不仅限于一般的日志分析，访问记录也是日志的一部分，除了计算网页浏览率以及访客人次外，电商的销售额、微博的评价、滞留时间等信息，只要由网络连结在一起的数据，Web 挖掘都可以做，也可整合线下及在线的数据库，实施更大规模的预测与推荐，毕竟凭借互联网的便利性与渗透力再配合网络行为的可追踪性与高互动特质，一对一的精确营销理念是最有可能在网络世界里完全实现的。"

听徐教授这么一说，大家更来兴致了，又是一翻七嘴八舌的讨论。

徐教授用手做了一个下压的手势，说："大家静一静，接下来，我就和大家一起重点探讨数据挖掘技术在互联网垂直搜索、电子商务、社交网络等方面的应用，垂直搜索用户行为侧重于 Web 链接结构和使用模式挖掘，面向电子商务和社交网络方面则侧重于 Web 内容挖掘和使用模式挖掘。"



9.2 垂直搜索引擎中的数据挖掘

徐教授："相信我们在座的很多领导，因为工作需要到外边出差。都会涉及飞机票/火车票以及酒店的预订问题。这个飞机票/火车票以及酒店的预订行为在互联网上的实现基础就是垂直搜索。"

"徐教授，搜索估计大家都知道百度，您说的垂直搜索我是第一次听说"，黄主任说出了自己心中的疑惑。

只见马处长站起来说："这个我知道，垂直搜索是针对某一个具体行业的专业搜索引擎，它是对搜索引擎的细分和延伸，垂直搜索的特点是更专、更精、更深，且具有特定行业的色彩。"

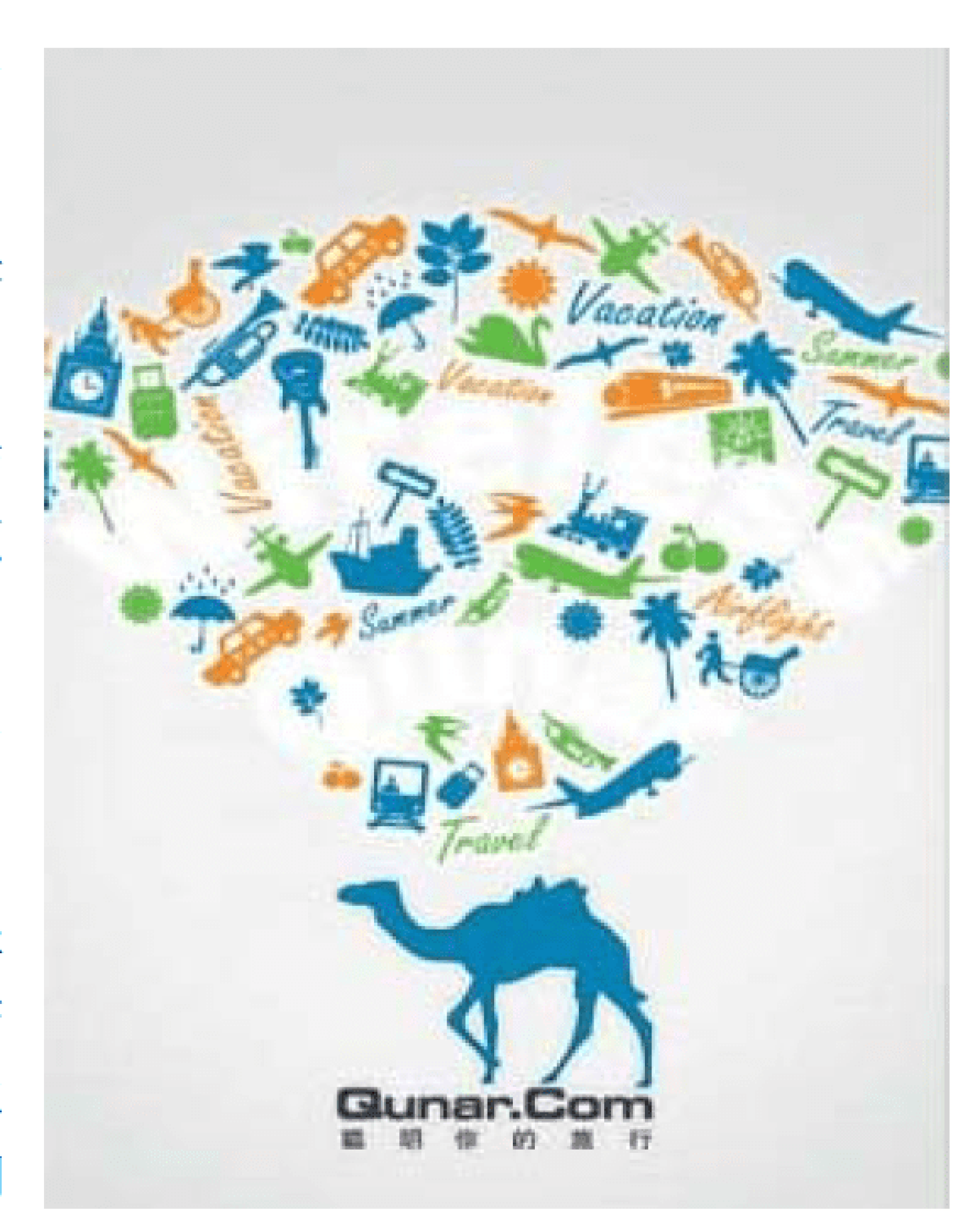
徐教授说:"马处长，你说的很对，随着互联网信息的不断增长，垂直搜索引擎成为互联网用户必不可少的助手。大部分垂直搜索引擎都在后台服务器详细记录了用户搜索的完整过程，包括用户的 IP 地址、搜索时间、输入的查询词、点击的 URL 等。通过对互联网用户使用搜索引擎的行为进行分析，可以挖掘用户的搜索规律，揭示用户的搜索意图，一方面可以提高搜索质量，满足用户的需求，另一方面可以改善和提高搜索引擎的性能，提高搜索引擎的知名度和扩大其市场份额。"

"有没有用户使用覆盖比较广、品牌熟悉度高的垂直搜索引擎？"台下的姚局长也好奇地问道。

李部长神气地说："说起垂直搜索引擎，就得说'去哪儿'，去哪儿从事旅游垂直搜索行业，为消费者提供国内外机票、酒店、火车票和旅游度假等专业搜索服务，帮助用户实时获取全方位旅游产品信息。就我所知，去哪儿旅游网可以搜索超过 10 万家酒店、2000 家专业机票、火车票以及度假产品的供应商网站。"

徐教授说："以酒店搜索为例，搜索引擎将酒店资料、在线的房间价格、行政区划和定位，以及用户评论等信息综合在一起，并通过数据挖掘等方法，以便让用户有更好的酒店查询效率和感受。"

徐教授说：使用搜索引擎面临许多问题，比如：



- 用户使用搜索引擎是为了快速找到自己需要的信息；
- 用户不能面对无限的搜索结果，搜索引擎提供的结果必须是很收敛的；
- 很多情况下，用户并不能恰当的描述自己需要什么；
- 即便使用搜索引擎，用户也经常要面对较多的无效信息，需要花费很多时间和精力去比较、去重复查找结果，这限制了搜索引擎的价值。

张处长说：“准确定位用户搜索行为确实有一定难度，以搜索包子为例，他可能想到的如：多大的包子好包、怎么蒸包子、包子的营养、纸壳馅包子、哪买的包子好吃，其实说不定他最想知道的是包子到底能不能喂狗”。

听了张处长的幽默话语，大家都乐成一片。

李主任继续思考说：“面对这类问题确实比较棘手，我能想到的解决问题的途径：从查询入手，搜索引擎帮助用户更好的描述自己需要什么，例如各种查询向导，用可视化的选择代替语言以及对自然语言的联想。利用扩散思维，比较用户找凳子，互联网提供方凳、圆凳、靠背、无靠背、三腿、四腿凳等的查询向导。”

姚局长作为武侠小说迷，提供了自己的见解：“从结果入手，搜索引擎从潜在的搜索结果中，总结出更有通用性、更可能产生价值的结果优先提供给用户。用户找无功秘籍，互联网提供葵花宝典、降龙十八掌、佛山无影脚、小李飞刀等。”



徐教授听后满意的点头道："垂直搜索引擎通常每天都会收到数百万用户提交的查询词，这些查询词对搜索引擎来说是非常有价值的，它们可以帮助搜索引擎进一步调整其检索与排序算法，从而给用户返回更好的查询结果。然而，由于查询词数目的巨大，搜索引擎不可能直接利用这些数据；另一方面，从查询词本身来说，一个特定的查询词是很随意的，无法清晰地表示一种用户信息需求。因此直接利用这些查询词对搜索引擎来说是没有太大意义的。为了解决这一问题，一个很直观的方法就是尝试对这些查询词进行聚类。在一个类别内部，一组查询词作为一个整体代表了一种用户信息需求或者是用户兴趣。"

学员们听了徐教授的话都被折服了，原来李主任和姚局长的简单想法被徐教授提升到了智能查询的高度。

徐教授接着说："保证用户能搜索到自己想要的，同时还要保证搜索的效率。以搜索酒店为例，评估酒店的查询效率有两方面：一方面用户消耗时间精力来检查酒店结果名单是成本，用户阅读感兴趣的酒店详情或预定房间，是产生价值；另一方面追求低成本、高价值，优化衡量用户查询效率的指标——结果页的酒店转化率。此外，单位时间内，酒店点击（或预定）次数/在搜索结果中的展示次数，即转化率，也是考察的一个指标。例如展示 50 次，点击 6 次，则转化率为 12%。"

李处长略微有点激动："转化率我听过，很重要。那怎么提升酒店转化率呢？"

徐教授回应道：提升酒店的转化率有以下几个方法：

- 显然，所有符合用户搜索条件的结果中，将转化率较高的那些酒店优先呈现给用户，是个提升查询效率的办法；
- 统计和排名酒店的转化率（在终极分析中，一切知识都是历史；在抽象的意义下，一切都是科学教学；在理性的基础上，所有的判断都是统计学），我们考虑如下因素：修正展示次数、重视统计精度、动态周期更新；
- 在此基础上，还需区别用户差异，实现定向推荐酒店。

姚局长一脸迷茫："徐教授，这些名词都比较专业，我都听糊涂了，修正展示次数？怎么个修正法儿呢？"

徐教授说："别着急，听我解释，观察用户在搜索结果页的点击行为，我们发现：酒店在搜索结果页出现的位置，在很大程度上影响了酒店被点击的概率。位于结果页第一位的酒店的点击数最多，占总点击数的百分之二十以上，前三位的点击额度占百分之四十以上，大体上位置越靠后其点击的数量就越少。我们来看一下对"展示次数"的修正：酒店在第 i 个位置上展示了一次，在逻辑上，我们认为此酒店被展示了 Ci 次，转化率的公式修正为：此酒店的被点击次数/Σ（在位置 i 的展示次数*Ci）。"

徐教授接着说："我们还要重视统计精度，有些酒店的展示次数很少，偶尔有一两个人点击，这时该酒店的转化率会非常高。酒店 A 展示了 3 次，被点击 2 次，转化率是 66%。酒店 B 展示了 10000 次，被点击 1000 次，转化率是 10%。如果我们简单的比较 66%和 10%这两个数值，会认为用户更喜欢酒店 A，但事实上，这样的做法忽视了酒店转化率的精度。如果将酒店的转化率 p 视为伯努利实验中事件发生的概率，将展示的次数 n 视为实验次数。那么实验的方差可以表示为：

$$\delta = \sqrt{p(1-p)/n}$$

为了限制方差的大小，也就是提高统计精度，要求方差与均值的比值满足下面的条件：

$$\delta/\mu = \delta/p \le p_1\%$$

其中 p1 为待确定的参数，取值范围为[1，30]。

通过控制置信度区间，我们可以有效地控制计算结果的可信程度，使得在适应实际应用与性能要求的条件下，满足统计的精确程度。"

大家没想到这背后有这么多学问，都听得更聚精会神了。

徐教授说："现实生活中，消费者的酒店偏好往往会随时间发生变化。造成这些变化的因素，有些是可以预见的，如随季节的周期性变化或大型活动、节日等。有些因素是不可预见的，如酒店打折促销、某地偶然的重大事件等。搜索引擎必须自动监测这种变化，动态周期更新，及时做出调整和响应。"

李部长说："这个好办，设定统计周期为固定时间周期，例如每星期作为一个周期，这样直观。"

张行长表达了不同的意见："嗯，这么做容易理解，但是我担心使用这种方法存在局限性。对于有众多搜索行为的特大城市，每星期作为一个统计周期，可能频率偏低，不能反映即时变化。"

姚局长插进来说：“是啊，对于鲜有人访问的小型城市，相同的每星期作为一次统计周期可能没有必要，因为访问量可能只有上百次的搜索和点击，采集不到足够的样本。”

大家都陷入了思考，那该怎么去平衡呢？

徐教授说：“嗯，大家想的都很周全，改进后的办法，第一种是固定访问次数周期。例如某城市每万次的搜索或点击，作为一次统计周期。这种方法在重要城市，因为搜索量大，满足一次统计周期条件的时间相对较短，在小型城市，因为访问量较小，满足一次统计周期条件的时间相对较长。如果在小型城市因为举办旅游节等活动，而吸引到大量的搜索，这种方法也会相应缩短一次统计周期的时间，从而实现自动调整。”

听了徐教授的话，黄主任意犹未尽：“徐教授，那还有其他什么改进方法呢？”

徐教授说：“另外一种比较经典的方法就是多周期加权统计，使用单周期的转化率统计得到的酒店排名，波动性会比较大。一方面，我们期望这种波动，它反映了用户喜好的即时变化。另一方面，我们还期望某些酒店有长期靠前的排名，它反映了这些酒店有现实中的竞争优势。因此，我们采用多个统计周期，使用转化率加权平均的最终计算办法，越近的周期权重越大，较好的均衡了两方面的考虑。”

李部长问道：“徐教授，之前您介绍的对酒店转化率的统计，所依据的是“用户整体”的行为统计，没有区分用户间的差异。”

徐教授赞许地点了点头。而下学员们陷入了思考，怎么区分用户间的差异性，收集每个用户对每家酒店的喜好显然不现实。

徐教授解释道：“我们可以对用户群体间的差异进行分析，实现针对特定用户的酒店展示或推荐。利用聚类分析有助于研究搜索引擎中的用户行为模式，为提高搜索引擎的检索效果提供支撑。”

黄主任说：“有的用户群无视价格，只住高档酒店。”

姚局长说：“有的用户群价格受得起，卫生最重要。”

李部长说："第二天一早见客户，酒店交通地点必须好。"

张行长说："也有手头太紧了，找个便宜酒店凑合住的群体。"



徐教授说："大家都说的很好，假定用户 A 和用户 B，他们拥有相近的价值取向和思维方式，他们在搜索引擎上会体现成类似的操作习惯，关注同样的酒店要素。用户 A 喜欢的酒店 i，用户 B 也有较大的概率会喜欢。和酒店 i 类似的酒店 j，也可能被用户 A 和 B 喜欢。"

李部长说："徐教授，我刚听到你讲相同类型的用户群关注同样的酒店要素，不知道这个要素都指的是什么呢？"

徐教授解释道：用户搜索酒店时的要素有：

- 一是对搜索引擎的操作习惯（区分用户群体），在使用介质上（PC 终端、手机等），在查询时间上（白天、晚间）等，在登入来源上（引擎跳转、直接网址等），在搜索的来源词上（旅游、酒店等）；
- 二是筛选酒店时的要素倾向性（区分用户群体、定位群体喜好），考虑房型、价格、设施、服务、周边交通等；
- 三是用户点击或预定酒店的要素倾向性（定位群体喜好）。

假设这些要素的集合是 I，那么表示用户集合 U 对各要素的倾向性，可以用矩阵表示。

听到这里，学员们心中的疑惑逐渐被解开。

徐教授接着说："对于任何酒店 i、j，可以得到酒店相似度的公式。对于任何酒店，都可以得到 k 个近邻，即数据挖掘里面的 K 近邻分类算法，可将相似度高的酒店放在一个类别中。通过一个在线用户的操作习惯、关注要素，我们即可预测评估这个用户对酒店的喜好程度。这样，我们得到一个酒店排名，即这个用户有可能选中的酒店。将这个酒店排名与前面得到的酒店转化率的排名进行混合，即为我们提供给用户的最终结果。"

姚局长感慨道："难怪酒店排名这么精准，像是电脑入侵我脑袋了似的。"

徐教授说道："实践是检验真理的唯一标准，将用户随机地分成 A、B 两组，使其分别看到两组来自不同算法的结果页。一组采用随机的酒店分布方式展示，一组用前面介绍的推荐方式展示，对比这两组用户的页面转化率（酒店点击或预定次数/结果页次数）。从结果上看，根据数据挖掘得到的推荐算法要明显好于随机算法。"

学员们都为徐教授严谨的治学态度折服。

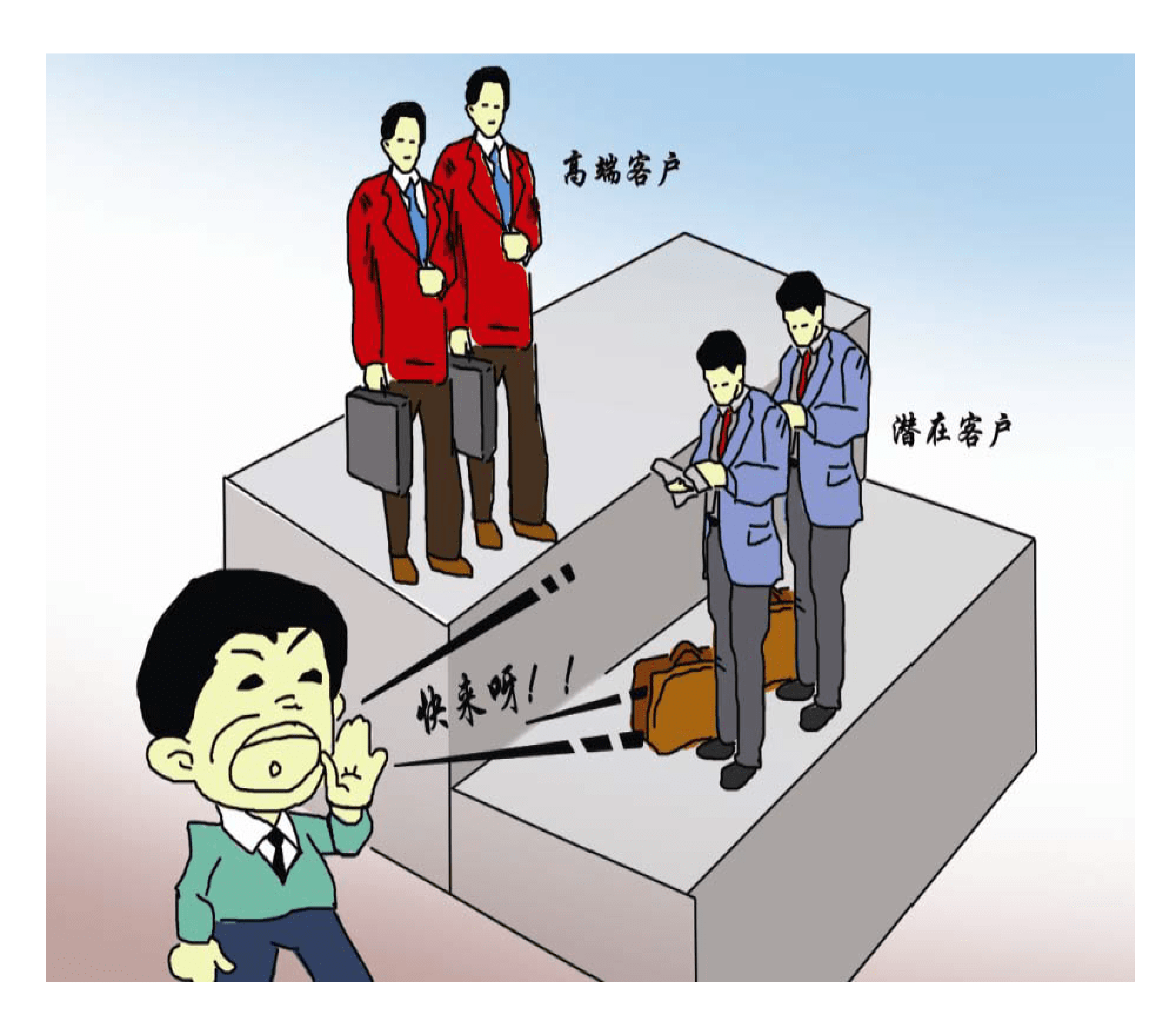
徐教授顿了顿："当然，AB 测试也不限于对比算法结果，也可用于比较算法本身的参数选择，实现结果页最好的展示结果。综上所述，通过对用户搜索酒店行为和结果进行的数据挖掘，我们提高了用户使用搜索引擎寻找酒店的效率，为用户带来更实际、更快捷的旅行便利。同样的数学方法，也使用在我们其他的业务领域如机票、火车票、度假、旅游指南，以及一起玩社区。"

听了徐教授对本节课程的总结，学员们都感慨万千，真是生活处处皆学问。

9.3 面向电子商务的数据挖掘

徐教授："首先让我们都明确一个概念：电子商务，是指在互联网上进行的商务活动，广义上不仅包括通过 Internet 买卖产品和提供服务，还包括企业内部和企业间的商务活动，不仅是硬件和软件的结合，更是把买家、卖家、厂家和合作伙伴在互联网上利用 Internet 技术与现有的系统结合起来开展的业务。"

姚局长说："嗯，我女儿大学专业就是电子商务，所以我对电子商务知道一点儿。比如淘宝、京东、苏宁易购、亚马逊等都是电子商务平台，只是对数据挖掘怎么在上面发挥作用还不了解"



徐教授："面向电子商务数据挖掘的任务主要表现在客户关系管理方面。由于互联网的存在，电子商务使企业和客户之间的交流更加方便、频繁和便捷，因此，企业更多的需求是如何通过电子商务的数据挖掘掌握更多客户的信息动态，以便改进企业与客户交流的方式和提出新的交流内容等。在留住老客户的同时也要善于挖掘新客户，利用分类技术可以在电子商务网站上找到潜在客户，通过挖掘 Web 日志记录，先对已经存在的访问者进行分类，然后从它的分类可以找到潜在的客户。"

李部长："徐教授，电子商务网站本身是通过搜索引擎提供网上交易，拿淘宝来说，这其中涉及到您上节提到的搜索背后的技术支撑，此外，面对几亿卖家和买家的海量信息交互，电子商务的数据挖掘必然还有一些其他特点。"

张行长："徐教授，根据之前的学习，我有这样一个想法，在电子商务中，客户聚类应该也是一个重要的方面。比如针对 Web 进行模式分析，挖掘出具有相似浏览模式的客户。然后，通过对具有相似浏览行为的客户进行分组，分析组中客户的共同特征，帮助电子商务的组织者更好的了解自己的客户，向客户提供更适合、更面向客户的服务。"

徐教授："是的，你们都说的很对。在前一节我们已经知道，用户搜索需求的分析和精确表示包含很多内容。通过 Web 内容挖掘，可进行电子商务海量商品的信息采集。针对电子商务网站的用户搜索，第一需要准确地把握用户搜索 query 分析，其次根据用户行为投放定向广告，比如我们可以在优酷页面上看见淘宝广告投放。通过 Web 数据挖掘，电子商务的经营者可以得到可靠的市场反馈信息，分析顾客的未来行为，有针对性的进行电子商务营销活动。根据产品的访问者的浏览模式来决定广告的位置，增强广告针对性，提高广告的投资回报率，从而降低运营成本，提高企业竞争力。另外一个特点就是商品的推荐系统应用，比如淘宝页面上的'推荐同类已购买产品'，以及在亚马逊页面上的'可能感兴趣的产品'等。进一步地，根据挖掘客户活动规律，有针对性的在电子商务平台下提供个性化的服务，比如针对不同的用户提供不同的服务策略和服务内容的服务模式，其实质就是以用户需求为中心的 Web 服务。它通过收集和分析用户信息来了解用户的兴趣和行为，进而实现主动推荐服务。因此，通过网络提供的个性化服务可以有效地解决用户信息过载和信息迷失的困境，还可以帮助企业建立友好的客户关系。因为电子商务本身是一个信息化非常完全的系统，所积累的数据通常存储在电子商务系统的数据库中，这些数据库一般是分布式的，而用户主要是从网络上获取这些数据，因此对电子商务使用的数据挖掘主要是分布式数据挖掘。"



黄主任：“徐教授，听你这么一说，我对数据挖掘在电子商务中的应用有了很大认识。”

徐教授进一步说道：“通过 Web 使用模式挖掘，可辅助商家理解用户行为，从而改进站点结构，调整销售策略，提供个性化服务。今天我们重点掌握 Web 使用模式挖掘。”

彭处长：“徐教授，推荐产品应用是不是也用到了 Web 使用模式挖掘呢？”

徐教授笑着说：“是的，一般地，面向电子商务的 Web 使用模式挖掘有以下模式可被发现：路径分析、关联规则挖掘。首先我们来掌握路径分析，路径分析可以用来发现 Web 站点中最经常被访问的路径，从而可以调整站点的结构。”

马处长：“有个例子说明就好了，我是碰见网站之类的分析就怵。”

李部长调侃道：“别怵，有徐教授呢，给你开个良方。”

徐教授看他俩打趣，接着道：“那就简单地举例说一说，比如观察某电子商务网站的路径分析，我们可以发现如下的信息：

- 除了主页，70%的客户是从/product/page1 进入网站的；
- 60%的客户是从/company/进入/company/page1 的，但是他们很少从/company/进入/company/page2 或/company/进入/company/page3；
- 85%的客户经过 4 级链接后，就会离开网站。”

李部长见马处长神情专注，便问："马处长，看出来这几条链接信息背后表达的含义了吗？给大伙说说"……

马处长也不慌，自信地说道：我只是发表一下我的想法，有不同意见希望大家批评指正。针对徐教授说的三条信息，我们分别来解读：

- 第一条链接信息说明/product/page1 对于用户来书最有用，可以在这页加入重要的超链接或者网站目录结构。
- 第二条链接信息说明/company/page1 包括很多有用的信息，但是用户很多不是直接进入该页面，而是需要通过其他链接进去。
- 第三条链接信息说明多数用户不愿意浏览链接超过 4 层的页面，所以最好将重要页面放在小于 4 层的位置。

听完马处长的话，大家才明白他之前说的都是谦虚话，要不然怎么可能说得连徐教授都不停点头肯定呢。

徐教授："我们再来看看关联规则在 Web 使用模式挖掘中的应用。关联规则主要用于事务数据库中关联知识的发现。比如用于电子商务，可以挖掘客户购买商品的模式：

Age（X，"20，……，29"）&income（X，"5k,……,10k"）=>buys(X,"computer"),[support=2%,Confidence=80%]表示所有用户中有 2%持度，年龄为 20 至 29 岁，月收入为 5000 至 10000，且购买计算机。这个年龄和收入组的用户购买计算机的可能性为 80%。电商企业挖掘出某些比较受客户欢迎的特征产品后，可能增强此类产品的设计和生产，以便将他们精确地推销地合适的用户。"

姚局长："徐教授，你说的这个让我想起了购物篮分析，只是这个消费行为是在互联网上而已，对吧？"

李部长："看来姚局长已经领会到关联规则的实质了，佩服。"

徐教授笑道："嗯，看来我之前讲的大家都掌握了。除了购买模式的挖掘，关联规则还可以寻找出被频繁访问的网页组，帮助我们了解它们之间的相互关系。40%的客户既访问了/company/page1，又访问了/company/page2；30%的客户在访问过/company/special 后，在/company/page1 中购买了商品。这样在进行产品推荐过程中，我们可以将加大这些页面的广告投入等费用。总体上来看，电子商务对数据结果的应用通常是针对电子商务系统的。"

张行长说："徐教授，是不是可以这么理解：电子商务数据挖掘的目的是提高企业竞争力，但是电子商务领域中的数据挖掘提高企业竞争力的方式通常是对电子商务系统的改进。比如给客户推出个性化页面，把用户最感兴趣的信息放在首页，或者像你上面说的例子中放在关联性比较强的页面，从而更能吸引用户。"

徐教授："是的，通过对客户的行为记录和反馈情况进行挖掘，为站点设计者提供改进的依据，从而站点设计者可以进一步优化网站组织结构来提高网站的点击率。利用关联规则，针对不同客户动态调整站点结构，使客户访问的有关联的文件之间的链接更直接，客户可以方便地访问到想要访问的页面，更具有便利性。同时提高站点质量，给客户留下好印象，增加下次访问的机率。另外，对网站上各种数据的统计分析有助于改进系统性能，增强系统安全性，并提供决策支持。"

黄主任："听了大家的话，我觉得很有启发。在电子商务中，虽然每个用户在不同的时期会有不同的访问模式，但其长期趋势是稳定的。因此通过分析一定时期内商务站点上的用户的访问信息，可以发现该站点潜在的客户群体、聚类客户、相关页面等，这些信息对于电子商务网站来说是非常有价值的。"

李部长附议道："对商品访问情况和销售情况进行挖掘，企业能够获取客户的访问规律，确定顾客消费的生命周期，根据市场的变化，针对不同的产品制定相应的营销策略。"

高处长也积极发言："电子商务跨越了时间、空间距离，客户可以自主选择销售商。而销售商通过挖掘客户访问信息，了解客户的浏览行为，根据客户的兴趣与需求，

向客户做动态地页面推荐和提供定制化的产品，提高客户满意度，延长客户驻留的时间，最终达到留住客户的目的。”

徐教授：通过这节课，我们可以概括出如下结论。整体而言，电子商务中的数据挖掘具有以下特性：

- 资料收集容易且不引人注意，所谓凡走过必留下痕迹，当访客进入网站后的一切浏览行为与历程都是可以立即被记录的；
- 以交互式个人化服务为终极目标，除了应不同访客呈现专属设计网页之外，不同的访客也会有不同的服务；
- 可整合外部来源数据让分析功能发挥地更深更广，除了 logfile、cookies、会员填表数据、在线调查数据、在线交易数据等由网络直接取得的资源外，结合实体世界累积时间更久、范围更广的资源，将使分析的结果更准确也更深入。

姚局长：“徐教授，听了这节课，受益匪浅。那未来电子商务的数据挖掘方向如何呢？”

徐教授：“利用数据挖掘技术建立更深入的访客数据剖析，并赖以架构精准的预测模式，以期呈现真正智能型个人化的网络服务，是互联网数据挖掘努力的方向——提供比我们自己更懂自己的网上交易服务。”

教室里学员人心振奋，都在憧憬未来电子商务给我们的生活带来的便利和美好前景。

9.4　社交网络中的数据挖掘

徐教授：“随着 Facebook 的上市，社交网络再次成为人们关注的焦点。社交网络，也就是网络+社交的意思。通过网络这一载体把人们连接起来，从而形成具有某一特点的团体。与传统的论坛、博客相比，社交网络是虚拟世界与现实世界的桥梁，在互联网上将现实生活中人与人之间的关系建立起来，Facebook、Twitter、LinkedIn 分别代表三种不同的社交网络。谁主动给大家介绍一下这三种社交网络所代表的类型？”



李部长当仁不让：“Facebook 是基于朋友之间强关系的社交网络，有助于朋友之间关系的维系和改善。Twitter 是基于单向关注的弱关系的社交网络，这样的网络有利于塑造意见领袖和消息的传播；而 LinkedIn 是面向商务人士的职业社交网络，帮助用户利用社交关系进行商务交流以及求职招聘。”

张行长：“前面说的都是国外的，针对国内社交网络表达一下我个人的理解，国内的腾讯

QQ、人人网类似于 facebook 偏向于朋友之间关系维护；微博类似于 Twitter，是一个基于用户关系的信息分享、传播以及获取平台，用户可以通过 Web、WAP 以及各种客户端组建个人社区，以 140 字左右的文字更新信息，并实现即时分享；猎聘网类似于 Linkedin 专注于高端招聘领域的社交关系改善。”

徐教授：“大家的认识都很独到，社交网络每天都会产生大量的用户数据，并且具有空前的规模性和群体性，吸引着无数研究者从无序的数据中发掘有价值的信息。这就像概率统计中经常举的‘投硬币算其正反面概率’的例子，从几次的投掷结果中很难看到规律，但通过几万次的大量投掷实验，便很容易看出正反面的出现次数几乎相等的规律。社交网络上产生了大量的规模化、群体化的数据，吸引了包括计算机科学、心理学、社会学、新闻传播学等领域专家和学者对其进行研究和探索，希望能够借助更强的社交网络的分析和处理能力发现更多人类尚未探索出的规律。首先我们来确认一点：为什么要分析社交网络数据？”

黄主任：我觉得从三个方面可以说明分析社交网络的重要性。

- 用户量在这，新浪微博、腾讯微博、人人、腾讯朋友、QQ 空间、开心 001 等这几个大平台注册用户加起来比中国人口还多；
- 用户停留时间在这，有数据显示用户，95%的社区网民平均花费在社区的时间要超过一个小时；
- 最关键是用户喜怒哀乐都在这，因为用户现实的朋友在这，用户不真实的朋友也在这，用户不认识但关注的人都在这。

总之社交网络就是用户真实生活的一个反映，或者说真实生活在社交网络就是人生活的一部分。

徐教授点头肯定说：在社交网络中，你不止是你，你是数据世界的一部分。或者你可以看看这个世界上正在发生什么，你，就是数据本身。社交数据主要包含很多讯息，比如：

- 用户社交关系链：6 度-（世界上任意 2 个人只需 6 个人就能建立联系）、250+（每一个人背后有至少 250 人的圈子）；
- 用户关注内容/人：你为什么关注李开复，而不关注凤姐？为什么关注“空”姐，而不关注梁咏琪？这些都反映出个性喜好；
- 用户在网络呈现的属性：属性也可以理解为角色，每个用户存在多种角色。这个属性越来越接近于人本身，体现在关注、粉丝、评论、转发，标签等。

那分析社交数据能让我们知道什么？

姚局长：我个人的一点想法，对于企业来说：

- 第一可以找到自己的客户，客户主要属于哪个社交平台、有什么人口特征（角色模型）、他们的购买倾向及使用倾向；
- 第二，获得品牌和客服信息，品牌舆情分析走势，自身品牌有效传播者（也许是凤姐、也许可能是互联网的那点事）；
- 第三，知悉竞争对手在干什么，竞争对手的影响策略，如果你哪里做不好，竞争对手会告诉你。

徐教授，就是不知道这些信息怎么分析获得了。

徐教授：利用计算机来处理社交网络往往会将整个社交网络看作是一个图的结构，每个用户就是图中的节点，人与人之间的关系就是节点之间的边，根据不同类型的社交网络，所构成的图可以是有向图也可以是无向图，关系的强弱也可以利用边上不同的权重来体现。有以下几种：

- 一是关联关键字跟踪，在微博不断跟踪某些关键字的变化，如产品名、品牌（假如产品名称老跟‘烂’、‘差’同时出现，就要提高警惕了）；
- 二是查看传播路径、引爆点。传播路径反映出产品和品牌的渗透力量，引爆点反映出潜在的价值传播点；

- 三是从批量用户中识别产品民间代言人。通过关键字、频率识别用户关注度的领域/兴趣，也可以统计用户所有粉丝的兴趣分布以及影响力。

我认为是所有行业，特别是受别人行为影响比较比较明显的行业，如电商、科技产品、电子产品等应该分析和加强关注社交数据。

彭处长：“徐教授，国外的社交网络数据挖掘研究都有哪些方面呢？”



徐教授：对于社交网络的分析和研究范围很广，也存在着许多有意思的研究和应用课题。例如，在社交网络中社区圈子的识别、社交网络中人物影响力的计算、基于社交网络信息的预测等。

- 第一，社交圈子的识别。社交网络最核心的就是人与人的关系，以及所形成的社交圈子，然而每个人根据自己的关系不同及兴趣不同可以属于多个社交圈子。在算法中以亲密度为首要指标和以兴趣为首要指标，也会得到不同的社交圈子划分。

- 第二，影响力的计算：在社交网络中，意见领袖因为其在网络上强大的影响力会对信息的传播，以及普通用户的行为造成巨大的影响。与现实社会一样，社交网络中的人也存在不同的阶级和不同的影响力。如何评价一个人在不同领域的影响力也是一个很重要的问题。

有学者提出了基于主题级别的影响力评价模型来尝试解决这个问题，该算法应用在大规模社交网络数据中显现出了较好的效果。

姚局长："这点我认同，对于每个人来说，其在不同领域的影响力也是不一样的。例如，李开复的影响力主要在科技领域，黄健翔的影响力在体育领域，薛蛮子的影响力主要在投资和公益的领域。"

徐教授：我们接着说。

- 第三，用数据预测未来也是社交网络的一个重要方向。

华尔街的多家对冲基金公司已经在利用 twitter 数据挖掘来衡量人们的情绪，发现公众的情绪数据与很多社会现象及事件具有很强的相关性，无论是'希望'的正面情绪，还是'害怕'的负面情绪的体现都预示着美国股市指数的下跌。在流行病预测方面，英国的科学家根据 Twitter 的数据来跟踪流感的爆发。他们主要基于用户发布信息中的关键词，例如'我头痛'等，并结合用户的发布地点，按区域与英国卫生部的官方数据进行比较，最终建立起一个预测模型。还有很多研究者也利用数据挖掘的方法对电影票房、美国大选的趋势和结果进行预测，并取得了令人惊喜的成果。

李部长："徐教授，通过社交网络数据的预测应用这么神奇，听着好像无所不能呢。"

徐教授："我们对于利用社交网络数据预测能力的态度也不能过于乐观，因为社交网络的预测是基于海量数据的，但目前对于海量文本数据的分析算法尚未达到理想的准确率。尤其对于'从文本信息来进行情绪判断'这个看似简单的问题，其本质是自然语言处理与情绪心理学的交叉问题。对文本情绪的判断也以基于词库及语法结构的判断和基于机器学习的方法为主。然而这些方法对于稍显复杂的、尤其是带有反讽

和隐含意的语言很难进行有效判断。此外，对于社交网络的使用群体不能完全代表有效的人群，因为使用社交网络的人群与年龄、地域、种族等方面都有很大差异，因此仅利用社交网络产生的数据进行预测很可能会与最终结果产生偏差，所以从人群角度进行科学有效的取样方法对于社交网络预测也是尤为重要的一个环节。”

马处长：“徐教授，能给我们大家举例介绍一下您带领的团队中从事的研究吗？”

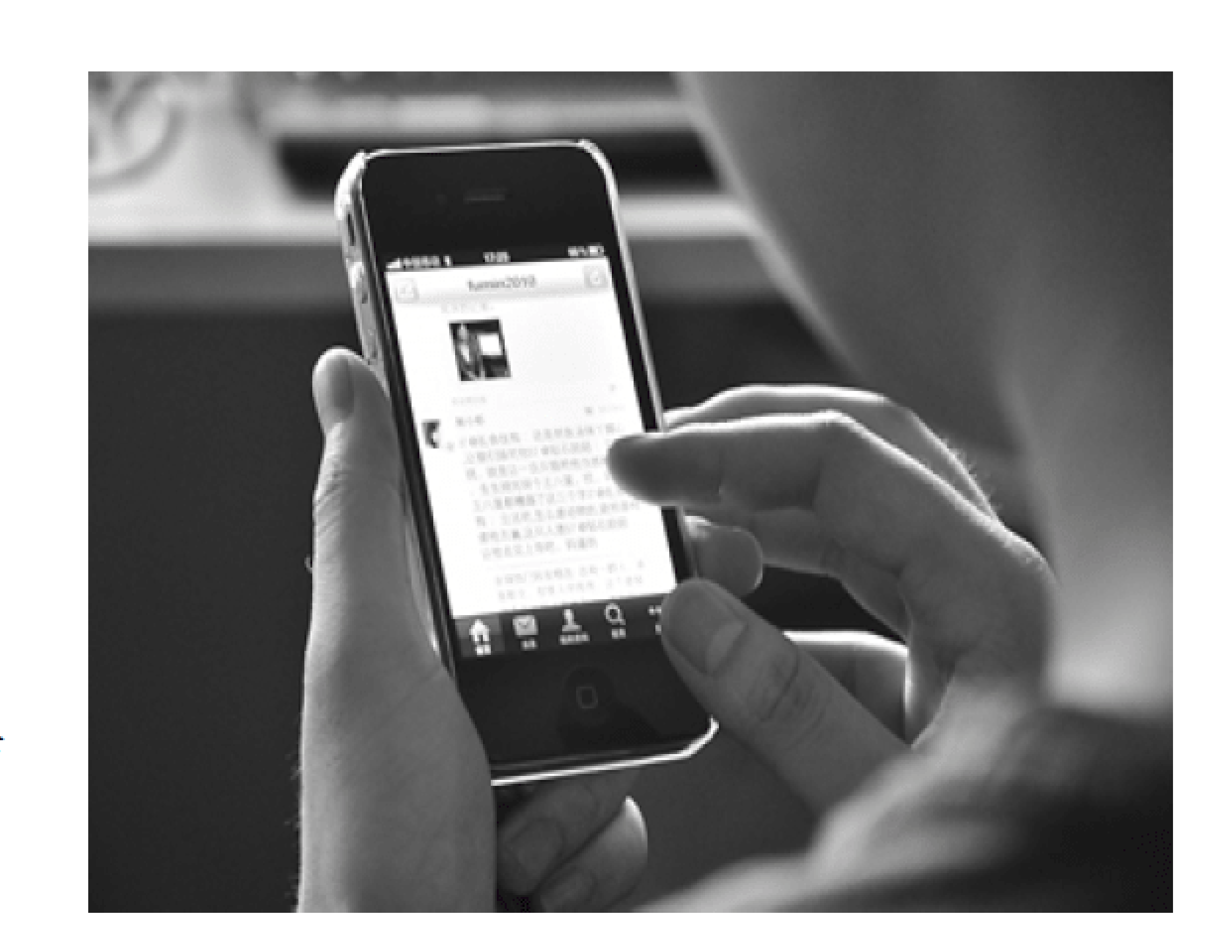
徐教授：我的研究团队中，有一个小组针对新浪微博的短链接进行了初步分析和研究。短链接，通俗来说，就是将长的 URL 网址，通过程序计算等方式，转换为简短的网址字符串。访问时，只要将原始网址与短链接对应，做映射，就可以实现跳转作用。各个 Web 网站推出自己的短链接无疑能在新浪微博等实时信息平台上占据更多优势：

- 一是提升品牌曝光率，让用户一眼就能知道链接出自哪里；
- 二是控制用户，基于上一点，用户同样也希望让好友知道自己分享的东西出自哪里；
- 三是整合并提高用户黏度，大部分有没有短链接都无所谓的服务一旦提供短链接及相应的配套广播功能，就会很轻而易举的留住用户；
- 四是完善自身提高可信度，提供短链接能让用户觉得该网站更可靠。

李部长：“参照微博现在火的程度，必定是很多企业必争之地。徐教授，赶紧给我们说说您的团队是怎么分析的呢？”

徐教授：从某时间段内的新浪微博数据中提取所有短链接，同时利用数据挖掘软件工具对这些链接进行了简单的分析和挖掘。从数据导入、数据清洗、数据变换到数据分析，主要进行新浪网站分析、频道分析、应用分析、游戏分析、团购网站分析、电子商务网站及微博关键词时间序列分析、最受欢迎歌手分析、最受欢迎歌曲分析的分类汇总，计算频数分布。通过研究发现微博信息传播网呈现小世界特征：平均最短路径很少接近 6，这与“六度分离”（世界上任意 2 个人只需 6 个人就能建立联系）理论不谋而合，发现微博信息传播网的度分布指数符合无尺度网络度分布，指数介于

2 和 3 这一特性。此外，我们另外一个小组针对社交网络的虚假账号和用户也进行了初步分析，对虚假用户的判断采用了以下 8 种用户行为特征：



- 博主的创建时间的一致性
- 博主的头像和名字
- 关注与粉丝比例
- 博主的粉丝质量
- 发布微博数量
- 最近 200 次转发的对象分布
- 转发同一条微博的频率
- 转发时所写的内容

针对以上 8 种特征，利用机器学习的分类算法训练模型，并利用模型进行后续虚假用户的预测，可以有效地发现虚假用户，在舆情分析中将其剔除，还原出真实的信息传播情况及舆情状态。

张行长：“听着这内容就估计里面的工作量很大呀！”

徐教授：“是的，我们对社交网络数据的认识和挖掘还处于相对初级的阶段，对这种大规模、高维度数据挖掘还在不断地演化。目前来看，文本语言的情感分析、社交网络的传播预测等很多基础性问题还不能得到有效解决，对深入研究社交网络造成了一些限制。但随着人工智能研究水平的不断提高，尤其是认知神经科学与人工智能技术相结合的研究，让我们看到了人工智能的新希望。当我们真正有能力解决这些问题以后，社交网络将会成为帮助我们预测未来趋势的有利工具。然而，充分使用社交网络数据也意味着暴露用户越来越多的隐私，因此，如何能够在用户隐私和数据完整中找到一个平衡点，也是今后数据工作者所要面临的问题。”

参考文献

【1】韩家炜.数据挖掘概念与技术[M].北京：机械工业出版社，2007

【2】Xu Fengmin，Zongben Xu，Honggang Xue. Sparse Index Tracking: An L1/2 Regularization Based Model and Solution [J].2011

【3】任学延，张代林，郑明东.炼焦配煤专家系统的研究与开发[J].燃料与化工. 2010，（11）：29-31

【4】张哲.基于支持向量机的变压器状态评估和故障诊断的研究.华北电力大学硕士论文库.北京.2009.22-36

【5】闰伟，刘云岗，王桂华等. 基于数据挖掘的交通流预测模型[J]. 系统工程理论与实践.北京.2010，（7）:1321-1324

【6】王亚琴.道路交通流数据挖掘.复旦大学.博士论文库.上海.2007.65-80

【7】郑俊华.基于支持向量机的高炉炉温预报的研究.浙江大学硕士论文库.杭州. 20-45

【8】贺诗波.自组织数据挖掘在高炉炉温预测控制中的应用.浙江大学硕士论文库.杭州.2008.30-55